ஒளிவழி எழுத்துணரியும் அதன் உருவாக்கமும்
(Optical Character Recognizer and its creation)

பேராசிரியர் இராசேந்திரன் சங்கரவேலாயுதன்
அமிர்தா விஷ்வ வித்யபீடம்,
கோயம்புத்தூர்
rajushush@gmail.com

கோயம்புத்தூர்

மார்ச்சு, 2023

ஆசிரியர் உரை

தமிழின் தொழிநுட்பவளர்ச்சி பற்றி நான் எழுதிவந்தபோது அதுகுறித்த பல கட்டுரைகள் எழுதி அவற்றை எல்லாம் என் மடிக்கணியில் உள்ளீடு செய்து வந்தேன். ஒளிவழி எழுத்துணரி (Optical Character Recognizer (OCR)) குறித்து நான் எழுதிய வரைவு சிலகாலமாகவே எனது மடிக்கணியில் உறங்கிக்கொண்டிருந்தது. அதற்கு உயிர்கொடுக்கும் எண்ணத்தில் இதைப் பதிவேற்றம் செய்கின்றேன். இதில் நான்கு இயல்கள் உள்ளன. விரிவாக எழுத வேண்டும் என எண்ணினேன். காலம் போதாமல் அம் முயற்சியைக் கைவிட்டு இந்நிலையில் ஒரு நூல் வடிவில் உங்கள் முன் சமர்ப்பிக்கின்றேன் .

அன்புடன்

சங்கரவேலாயுதம் இராசேந்திரன்

இயல்	உள்ளடக்கம்	பக்கம்
1	ஒளிவழி எழுத்துணாரியும் அதன் உருவாக்கமும்	*4*
2.	தமிழில் எழுத்துணரி	*57*
3.	தமிழுக்கு கூகுள் எழுத்துணரி தொழில் நுட்பம்	*83*
4	தமிழ் கையெழுதுப் படிவத்திற்கான எழுத்துணரித் தொழில்நுட்பம்	*100*

1. ஒளிவழி எழுத்துணரியும் அதன் உருவாக்கமும்

வாசிப்பு போன்ற மனிதச் செயல்பாடுகளில் இயந்திரத்தின் உதவி ஒரு பழைய கனவு. இருப்பினும், கடந்த ஐந்து தசாப்தங்களாக, இயந்திர வாசிப்பு (machine reading) கனவில் இருந்து நினைவுக்கு வந்துள்ளது. அமைப்பொழுங்கு அறிதல் (pattern recognition) மற்றும் செயற்கை நுண்ணறிவு (artificial intelligence) துறையில் தொழில் நுட்பத்தின் மிக வெற்றிகரமான பயன்பாடுகளில் ஒன்றாக எழுத்துணரி மாறிவிட்டது. எழுத்துணரியை உருவாக்குவதற்கான பல வணிக அமைப்புகள் பலவகையான பயன்பாடுகளுக்கு இருக்கின்றன; இருப்பினும் இயந்திரங்கள் இன்னும் மனித வாசிப்பு திறன்களுடன் போட்டியிட முடியவில்லை. எழுத்துணரி தானாக அடையாளம் காணும் செயல்படும் நுட்பங்களின் குடும்பத்திற்குச் சொந்தமானது.

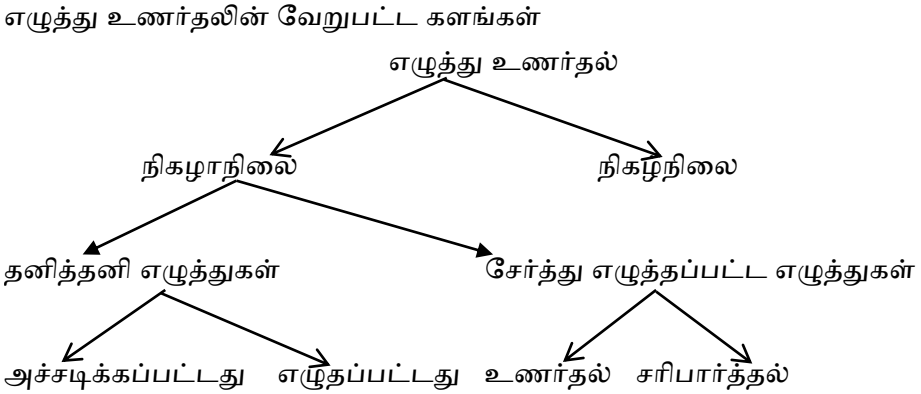
தானியங்கி அடையாளம் காணல்

கணினியில் தரவை உள்ளிடுவதற்கான பாரம்பரிய வழி விசைப்பலகை வழியாகும். இருப்பினும், இது எப்போதும் சிறந்த அல்லது திறமையான தீர்வாக இருக்காது. பல சந்தர்ப்பங்களில் தானியங்கி அடையாளம் காணல் ஒரு மாற்றாக இருக்கலாம். தானியங்கி அடையாளங் காணலுக்கான பல்வேறு தொழில்நுட்பங்கள் உள்ளன, மேலும் அவை பயன்பாட்டின் வெவ்வேறு பகுதிகளுக்கான தேவைகளைப் பூர்த்தி செய்கின்றன. எழுத்துணரி ஒரு தானியங்கி அடையாளம் காணும் கருவியாகும்.

அச்சிடப்பட்ட பனுவல்களை அப்படியே கணினிக்கு எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியை எழுத்துணரி என்பர். ஒளிவழி எழுத்து உணர்தல் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) (Optical character recognition) அல்லது ஒளிவழி

எழுத்துகளைப் படித்தல் (optical character reader (OCR)) என்பது தட்டச்சு செய்யப்பட்ட, கையால் எழுதப்பட்ட அல்லது அச்சிடப்பட்ட உரையின் படங்களை இயந்திர குறியீட்டு உரையாக மாற்றும் மின்னணு அல்லது இயந்திர மாற்றமாகும்.

ஒளிவழி எழுத்து உணர்தல் (Optical Character Recognition) ஒளியியல் செயலாக்கப்பட்ட எழுத்துக்களை அறிவதில் உள்ள சிக்கலைக் கையாள்கிறது. எழுத்து அல்லது அச்சிடுதல் முடிந்தபின், ஒளிவழி அறிதல் நிகழாநிலையில் (ஆஃப்லைனில்) செய்யப்படுகிறது, நிகழ்நிலை (ஆன்லைன்) உணர்தலுக்கு மாறாக, கணினி எழுத்துக்களை வரையும்போது அவற்றை உணர்கிறது. கையால் எழுதப்பட்ட மற்றும் அச்சிடப்பட்ட எழுத்துக்கள் இரண்டும் உணரப்படலாம்; ஆனால் செயல்திறன் நேரடியாக உள்ளீட்டு ஆவணங்களின் தரத்தைப் பொறுத்து அமையும்.



உள்ளீடு மிகவும் கட்டுப்படுத்தப்பட்டால், எழுத்துணரி அமைப்பின் செயல்திறன் சிறப்பாக இருக்கும்; இருப்பினும், முற்றிலும் கட்டுப்படுத்தப்படாத கையெழுத்துக்கு வரும்போது, எழுத்துணரி இயந்திரங்கள் வாசிப்பதற்கும் மனிதர்களுக்கும் இன்னும் நீண்ட தூரம். இருப்பினும், கணினி வேகமாகப் படிக்கிறது மற்றும் தொழில்நுட்ப முன்னேற்றங்கள்

தொடர்ந்து தொழில்நுட்பத்தை அதன் இலட்சியத்திற்கு நெருக்கமாக கொண்டு வருகின்றன.

மின்னணு ஆவணங்களிலிருந்து உரையை அடையாளம் காணவும் தேடவும் அல்லது உரையை ஒரு இணையதளத்தில் வெளியிடவும் இது பரவலாகப் பயன்படுத்தப்படுகிறது. இக்கருவி தட்டச்சு முறைப்படி சாவிப்பலகை வழி பனுவல்களை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது. ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டுப் பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய கருவியை உருவாக்குவது கடினமான செயலாகும். இருப்பினும் தமிழுக்கு இத்தகைய எழுத்துணரி உருவாக்கும் முயற்சிகள் எடுக்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

எழுத்துணரித் தொழில்நுட்பத்தின் முக்கிய நன்மைகள் நேரம் மிச்சப்படுத்தப்படுவது, பிழைகள் குறைதல் மற்றும் குறைக்கப்பட்ட முயற்சி. ZIP கோப்புகளில் சுருக்குதல், முக்கிய வார்த்தைகளை முன்னிலைப்படுத்துதல், ஒரு வலைத்தளத்துடன் இணைத்தல் மற்றும் மின்னஞ்சலை இணைப்பது போன்ற இயற்பியல் நகல்களுடன் திறன் இல்லாத செயல்களையும் இது செயல்படுத்துகிறது.

ஆவணங்களின் படங்களை எடுப்பது அவற்றை மின்னிலக்க (டிஜிட்டல்/digital) காப்பகப்படுத்த உதவுகிறது, எழுத்துணரி அந்த ஆவணங்களைத் திருத்தவும் தேடவும் கூடுதல் செயல்பாட்டை வழங்குகிறது.

ஒளிவழி எழுத்துணர்தல் (OCR) என்பது தரவு உள்ளீட்டு முறைகளில் மிகவும் பரவலாகச் செயல்படுத்தப்பட்ட வகைகளில் ஒன்றாகும். ஒளிவழி எழுத்துணர்தல் ஆனது மின்னிலக்க வருடல் (டிஜிட்டல் ஸ்கேனிங்/digital scanning) மற்றும் எழுதப்பட்ட அல்லது அச்சிடப்பட்ட உரையை உணர்தல் ஆகியவற்றை உள்ளடக்கியது. இங்கே உரை முதலில் ஃபோட்டோஸ்கான் செய்யப்பட்டு, பகுப்பாய்வு செய்யப்பட்டு, இறுதியாக எழுத்துக்குறி

குறியீடுகளாக மொழிபெயர்க்கப்பட்டுள்ளது. இயந்திர-குறியிடப்பட்ட இந்த உரையை எளிதில் தேடலாம் மற்றும் மின்னணு முறையில் திருத்தலாம்.

தரவு உள்ளீட்டுச் செயல்முறையை எழுத்துணரி பெரிதும் மேம்படுத்தியுள்ளது. இந்த மென்பொருள் கருவி, வருடல் (ஸ்கேன்/scan) செய்யப்பட்ட ஆவணங்களை விரைவாகத் தேடக்கூடிய உரைக் கோப்புகளாக மாற்ற உதவுகிறது. இன்று, ஆவணங்களை வருடல் (ஸ்கேன்/Scan) செய்ய வேண்டிய தேவை தொடர்ந்து அதிகரித்து வருகிறது, ஏனெனில் இது தேவைப்படும் போது இந்த ஆவணங்களை வசதியாகப் பார்க்க உதவுகிறது. வருடல் செய்யப்பட்ட ஆவணங்களையும் மின்னணு ஊடகம் மூலம் எளிதாகப் பகிரலாம்.

எழுத்துணரியின் பயன்பாடுகள்

எழுத்துணரியைப் பல்வேறு பயன்பாடுகளுக்குப் பயன்படுத்தலாம், அவற்றுள் சில:

- மைக்ரோசாஃப்ட் வேர்ட் அல்லது கூகிள் டாக்ஸ் போன்ற சொல் செயலிகளுடன் திருத்தக்கூடிய பதிப்புகளில் அச்சிடப்பட்ட ஆவணங்களை வருடல் (Scan-ஸ்கேன்) செய்கிறது.
- தேடுபொறிகளுக்கான அச்சுப் பொருளைக் குறிக்கிறது.
- தரவு உள்ளீடு, பிரித்தெடுத்தல் மற்றும் செயலாக்கத்தைத் தானியக்கமாக்குகிறது.
- பார்வைக் குறைபாடுள்ள அல்லது பார்வையற்ற பயனர்களுக்கு உரக்கப் படிக்கக்கூடிய ஆவணங்களை உரையில் புரிந்துகொள்வது.
- செய்தித்தாள்கள், பத்திரிகைகள் அல்லது தொலைபேசி புத்தகங்கள் போன்ற வரலாற்றுத் தகவல்களைத் தேடக்கூடிய வடிவங்களில் காப்பகப்படுத்துகிறது.
- வங்கி செல்லத் தேவை இல்லாமல் காசோலைகளை மின்னணு முறையில் வங்கியில் இட உதவுகிறது.

- முக்கியமான, கையொப்பமிடப்பட்ட சட்ட ஆவணங்களை மின்னணு தரவுத்தளத்தில் வைக்க உதவுகிறது.
- புகைப்பட கருவி (Camera/கேமரா) அல்லது மென்பொருளைக் கொண்டு உரிமத் தகடுகள் போன்ற உரையை உணர முடிகிறது.
- அஞ்சல் விநியோகத்திற்கான கடிதங்களை வரிசைப்படுத்த முடிகிறது.
- ஒரு படத்திற்குள் உள்ள சொற்களை ஒரு குறிப்பிட்ட மொழியில் மொழிபெயர்க்க முடிகிறது.
- ஒளிவழி எழுத்து உணர்தல் (OCR) மேலும் பல நன்மைகளைக் கொண்டுள்ளது.

எழுத்துணரி அடிப்படையிலான தரவு உள்ளீட்டின் நன்மைகள்

எழுத்துணரியின் ஏராளமான நன்மைகள் இருந்தாலும், இது முக்கியமாக வணிகத்தின் செயல்திறனை அதிகரிக்க வணிகநிறுவனங்களுக்கு உதவுகிறது. மகத்தான உள்ளடக்கத்தின் மூலம் விரைவாகத் தேடுவதற்கான அதன் திறன் மிகவும் உதவியாக இருக்கும், குறிப்பாக அலுவலக அமைப்புகளில், அதிக ஆவண வரத்து மற்றும் அதிக அளவு வருடல் (Scan/ஸ்கேன்) ஆகியவற்றைக் கையாளும். எழுத்துணரி தரவு உள்ளீட்டின் சில முக்கிய நன்மைகள் பின்வருமாறு:

1) அதிக உற்பத்தித்திறன்

தேவைப்படும் போது விரைவாகத் தரவு மீட்டெடுப்பதை எளிதாக்குவதன் மூலம் வணிகங்களுக்கு அதிக உற்பத்தித்திறனை அடைய எழுத்துணரி மென்பொருள் உதவுகிறது. தொடர்புடைய தரவுகளைப் பெறுவதற்கு ஊழியர்கள் செலவழிக்க வேண்டிய நேரம் மற்றும் முயற்சி இப்போது முக்கிய நடவடிக்கைகளில் கவனம் செலுத்த ஆற்றுவிக்கப்படலாம். தவிர, ஊழியர்கள் தேவையான ஆவணங்களை அணுக மத்தியப் பதிவு அறைக்கு ஏராளமான பயணங்களை மேற்கொள்ள வேண்டியதில்லை, ஏனெனில் அவர்கள் தங்கள் மேசைகளிலிருந்து எழுந்திருக்காமல் அவற்றை அணுக முடியும்.

2) செலவு குறைப்பு

எழுத்துணரியைத் தேர்ந்தெடுப்பது, தரவு பிரித்தெடுப்பதை மேற்கொள்ள நிபுணர்களைப் பணியமர்த்துவதைக் குறைக்க வணிகங்களுக்கு உதவும், இது எழுத்துணரி தரவு உள்ளிடும் முறைகளின் மிக முக்கியமான நன்மைகளில் ஒன்றாகும். நகலெடுத்தல், அச்சிடுதல், கப்பல் போக்குவரத்து போன்ற பல்வேறு செலவுகளை குறைக்க இந்தக் கருவி உதவுகிறது. எழுத்துணரி தவறாக வைக்கப்படும் அல்லது இழக்கும் ஆவணங்களின் சிக்கலை நீக்குகிறது. மீட்டெடுக்கப்பட்ட அலுவலக இடத்தின் வடிவத்தில் அதிக சேமிப்பு இடங்களை தருகிறது. காகித ஆவணங்களைச் சேமிக்கப் பயன்படும் இடத்தை மிச்சப்படுத்துகின்றது.

3) உயர் துல்லியம்

தரவு உள்ளீட்டின் முக்கிய சவால்களில் ஒன்று பிழை அல்லது தவறு செய்வது. எழுத்துணரி தரவு உள்ளீடு போன்ற தானியங்கு தரவு உள்ளீட்டுக் கருவிகள் குறைவான பிழைகள் மற்றும் தவறான தன்மைகளை விளைவிக்கின்றன, இதன் விளைவாகத் திறமையான தரவு உள்ளீடு சாத்தியமாகிறது. இது தவிர, தரவு இழப்பு போன்ற சிக்கல்களையும் எழுத்துணரி வழி தரவு உள்ளீடு செய்வதன் மூலம் வெற்றிகரமாக சமாளிக்க முடியும். இதில் மனித சக்தி தவிர்க்கப்படுவதால், தவறான தகவல்களைத் தற்செயலாக அல்லது வேறுவழியில் உள்ளிடுவது போன்ற சிக்கல்கள் அகற்றப்படுகின்றன.

4) அதிகரித்த சேமிப்பு இடம்

நிறுவன அளவிலான காகித ஆவணங்களை எழுத்துணரியால் வருடி தகவல்களை ஆவணப்படுத்தலாம் மற்றும் பட்டியலிடலாம். இதன் அர்த்தம் என்னவென்றால் காகிதத்

தரவுகளை இப்போது மின்னணு வடிவத்தில் சேவையகங்களில் சேமிக்க இயலும்; இதனால் பெரிய காகிதக் கோப்புகளைப் பராமரிப்பதற்கான தேவை நீக்குகிறது. இந்த வழியில் எழுத்துணரி தரவு உள்ளீடு நிறுவனம் முழுவதும் "காகிதமற்ற" அணுகுமுறையை செயல்படுத்த சிறந்த கருவிகளில் ஒன்றாக செயல்படுகிறது.

எழுத்துணரி எவ்வாறு பயன்படுத்தப்படலாம்

1) தரவுச் செயலாக்கம் (Data Processing)

ஒரு பிரபலமான எழுத்துணரி பயன்பாடு தரவு உள்ளிடல் ஆகும். சட்ட மற்றும் பிற வணிக ஆவணங்களின் கடின நகல்களை பிடிஎப் (PDF) கோப்புகளாக மாற்ற நிறுவனங்கள் இந்த மென்பொருளை நம்பியுள்ளன, இதனால் ஊழியர்கள் ஒரு சொல் செயலி ஆவணத்தைப் போலவே உள்ளடக்கத்தைத் திருத்தலாம், வடிவமைக்கலாம் மற்றும் தேடலாம்.

2) தரவு வகைப்பாடு (Data Classification)

தரவு வகைப்பாட்டிற்கு எழுத்துணரியைப் பயன்படுத்தலாம்; வங்கிகள் காசோலைகளை மின்னணு முறையில் வைப்பு சேமிப்பு (deposit) செய்யவும் அஞ்சல் அலுவலகங்கள் கடிதங்களை வரிசைப்படுத்தவும் இத்தொழில்நுட்பத்தைப் பயன்படுத்த அனுமதிக்கிறன.

3) மற்றவை

ஒரு தரவுத்தளத்தில் சான்றளிக்கப்பட்ட சட்ட ஆவணங்களைச் சேர்க்கவும், தேடுபொறிகளுக்கான குறியீட்டு அச்சுப்பொறிகளுக்கும், பார்வைக் குறைபாடுள்ளவர்களுக்கு ஆவணங்களை உரையாக மாற்றவும் எழுத்துணரியைப் பயன்படுத்தலாம். மொழிபெயர்ப்புப் பயன்பாடுகள், கூகிள் புத்தகங்கள் போன்ற

நிகழ்நிலை (ஆன்லைன்) உரை தரவுத்தளங்கள் மற்றும் உரிமத் தகடுகளை உணரும்/அறியும் பாதுகாப்புப் புகைப்படக் கருவிகள் (கேமராக்கள்/Cameras) ஆகியவற்றிலும் எழுத்துணரி பங்களிப்பு செய்கின்றது.

4) செயற்கை அறிவுடன் எழுத்துணரியை விரிவுபடுத்துதல்

கடந்த ஆண்டு, உடைந்த அல்லது பகுதி எழுத்துக்களுக்கான செயற்கை அறிவு அடிப்படையிலான எழுத்துணரி 2019 ஜப்பானிய கலாச்சாரம் மற்றும் செயற்கை அறிவு சிம்போசியத்தில் அறிமுகமானது. ஜப்பானிய கர்சீவின் ஒரு வடிவமான குசுஷிஜியைப் படிக்க இந்த தொழில்நுட்பம் மிகவும் உதவியாக இருக்கும்; குசுஷிஜி இன்று ஜப்பானிய மக்கள் எழுதுவதில் இருந்து மிகவும் வித்தியாசமானது; சிலர் அதை துல்லியமாக படிக்க முடியும்.

செயற்கை அறிவு இயக்கப்படும் பயன்பாடுகளும் அதிகரித்து வருகின்றன. திறன்பேசியின் புகைபடக்கருவி (ஸ்மார்ட்போன் கேமரா) மூலம் பணிபுரியும் இந்த பயன்பாடுகள் வணிக அட்டைகள் மற்றும் பிற ஆவணங்களை வருடி, தாமாகவே தரவை மேகக்கணிக்கு ஏற்றும்; பின்னர் மக்கள் அதை தனிநபர் கணினி அல்லது தொலைபேசி மூலம் அணுகலாம்.

ஒலியால் வருடி எழுத்துக்களைப் உணர்ந்துகொள்ளும் ஒழுங்குமுறைகள்

ஒலியால் வருடி எழுத்துக்களை உணர்ந்துக்கொள்ளும் (ஒளிவழி எழுத்து உணர்வான்) எழுத்துணரி அல்லது ((ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்/ Optical character recognition or ஆப்டிகல் கேரக்டர் ரீடர்/optical character reader (OCR/ஓ.சி.ஆர்)) என்பது தட்டச்சு செய்யப்பட்ட, கையால் எழுதப்பட்ட அல்லது அச்சிடப்பட்ட உரையின்

படங்களை இயந்திரக் குறியீட்டு உரையாக மாற்றும் மின்னணு அல்லது இயந்திர மாற்றமாகும். இது வருடப்பட்ட ஆவணத்திலிருந்து, ஆவணத்தின் புகைப்படம், காட்சிப்புகைப்படம் (எடுத்துக்காட்டாக குறியீடுகளிலான உரை, ஒரு நிலப்பரப்பு புகைப்படத்தில் உள்ள விளம்பர பலகை) அல்லது ஒரு படத்தின் எழுத்தப்பட்ட துணைத்தலைப்பு உரை (எடுத்துக்காட்டாக: ஒரு தொலைக்காட்சி ஒளிபரப்பிலிருந்து) என்பனவற்றைப் பெற உதவும்.

பாஸ்போர்ட் ஆவணங்கள், விலைப்பட்டியல்கள், வங்கி அறிக்கைகள், கணினிமயமாக்கப்பட்ட ரசீதுகள், வணிக அட்டைகள், அஞ்சல்கள், நிலையான தரவின் அச்சுநகல்கள் அல்லது பொருத்தமான ஆவணங்கள் எதுவாக இருந்தாலும் - அச்சிடப்பட்ட காகிதத் தரவுப் பதிவுகளிலிருந்து தரவு உள்ளீட்டு வடிவமாகப் பரவலாகப் பயன்படுத்தப்படுகிறது; இது அச்சிடப்பட்ட உரைகளை மின்னிலக்க (டிஜிட்டல்) மயமாக்குவதற்கான பொதுவான முறையாகும்; மின்னிலக்க (டிஜிட்டல்) மயமாக்கப்பட்ட உரைகளை மின்னணு முறையில் திருத்த இயலும், தேட இயலும், மேலும் சுருக்கமாகச் சேமிக்க இயலும், நிகழ்நிலையில் பெற (ஆன்லைனில்) இயலும்; மேலும் புலனறிவுக் கணினியாக்கம், இயந்திர மொழிபெயர்ப்பு, (பிரித்தெடுக்கப்பட்ட) உரையிலிருந்து பேச்சு, முக்கிய தரவு மற்றும் உரையை ஆழ்ந்தெடுத்தல் போன்ற இயந்திர செயல்முறைகளிலும் எழுத்துணரி பயன்படுத்தப்படுகின்றது. ஒளிமூலம் எழுத்து உணர்வான் என்பது மாதிரி உணர்தல், செயற்கை நுண்ணறிவு மற்றும் கணினிப் பார்வை ஆகியவற்றில் ஆராய்ச்சித் துறையாகும்.

ஆரம்பகால பதிப்புகள் ஒவ்வொரு எழுத்துக்களின் படங்களுடன் பயிற்சியளிக்கப்பட வேண்டும், மேலும் ஒரு நேரத்தில் ஒரு எழுத்துருவில் வேலை செய்ய

வேண்டும். பெரும்பாலான எழுத்துருக்களுக்கு அதிக அளவிலான புரிதல் துல்லியத்தை உருவாக்கும் திறன் கொண்ட மேம்பட்ட அமைப்புகள் இப்போது பொதுவானவை, மேலும் பலவிதமான மினிலக்க/டிஜிட்டல் படக் கோப்பு வடிவ உள்ளீடுகளுக்கான ஆதரவுடன். சில அமைப்புகள் வடிவமைக்கப்பட்ட வெளியீட்டை மீள் உருவாக்கம் செய்யும் திறன் கொண்டவை; அவை படங்கள், நெடுவரிசைகள் மற்றும் பிற உரை அல்லாத கூறுகள் உள்ளிட்ட அசல் பக்கத்தை நெருக்கமாக மதிப்பிடுகின்றன.

முறைப்படி, எழுத்து உணர்தல் என்பது அமைப்பொழுங்கு உணர்தல் களத்தின் (pattern recognition area) துணைக்குழு ஆகும். இருப்பினும், எழுத்து உணர்தல் தான் அமைப்பொங்கு உணர்தலையும் படப் பகுப்பாய்வையும் (image analysis) அறிவியலின் முதிர்ச்சியடைந்த துறைகளை உருவாக்குவதற்கான சலுகைகளை வழங்கியது.

1) முதல் முயற்சிகள்

இயந்திரங்களால் மனித செயல்பாடுகளைப் பிரதிபலிப்பது, வாசிப்பு போன்ற பணிகளை இயந்திரம் செய்ய வைப்பது ஒரு பழங்கால கனவு. எழுத்து உணர்தலின் தோற்றம் உண்மையில் ஆயிரத்து எழுபதாம் ஆண்டில் காணப்படுகிறது. போஸ்டன் மாசசூசெட்ஸின் (Boston Massachusetts) சி.ஆர். கேரி (C.R.Carey) விழித்திரை ஸ்கேனரைக் (retina scanner) கண்டுபிடித்த ஆண்டு இது, இது ஒளிச்சேர்க்கைகளின் மொசைக் பயன்படுத்தி பட பரிமாற்ற முறையாகும். இரண்டு தசாப்தங்களுக்குப் பிறகு போலந்து பி. நிப்கோ தொடர்ச்சியான Polish (P. Nipkow) வருடியை/ஸ்கேனரைக் கண்டுபிடித்தார், இது நவீனத் தொலைக்காட்சி மற்றும் வாசிப்பு இயந்திரங்களுக்கு ஒரு பெரிய திருப்புமுனையாக அமைந்தது.

பத்னொன்பதாம் நூற்றாண்டின் முதல் தசாப்தங்களில் எழுத்துணரி உடனான சோதனைகள் மூலம் பார்வையற்றவர்களுக்கு உதவ சாதனங்களை உருவாக்க பல முயற்சிகள் மேற்கொள்ளப்பட்டன. இருப்பினும், மின்னிலக்க கணினியின் வளர்ச்சியுடன் ஆயிரத்து நாற்பதுகளின் நடுப்பகுதி வரை எழுத்துணரியின் நவீனப் பதிப்பு தோன்றவில்லை. அப்போதிருந்து வளர்ச்சிக்கான உந்துதல் வணிக உலகில் சாத்தியமான பயன்பாடுகள் ஆகும்.

2) எழுத்துணரியின் தொடக்கம்

1950 வாக்கில் தொழில்நுட்பப் புரட்சி அதிவேகமாக முன்னேறி வந்தது, மின்னணுத் தரவுச் செயலாக்கம் ஒரு முக்கியமான துறையாக மாறியது. தரவு உள்ளீடு பஞ்ச் கார்டுகள் மூலம் நிகழ்த்தப்பட்டது; அதிகரித்து வரும் தரவைக் கையாள்வதற்கான செலவு குறைந்த வழி தேவைப்பட்டது. அதே நேரத்தில் இயந்திர வாசிப்புக்கான தொழில்நுட்பம் பயன்பாட்டிற்குப் போதுமான முதிர்ச்சியடைந்து வந்தது; ஆயிரத்து தொள்ளயிரத்து ஐம்பதுகளின் (1950-59) நடுப்பகுதியில் எழுத்துணரும் இயந்திரங்கள் வணிக ரீதியாக மாறி கிடைத்தது.

முதல் உண்மையான எழுத்துணரி வாசிப்பு இயந்திரம் ஆயிரத்து தொள்ளாயிரது ஐம்பதில் ரீடர்ஸ் டைஜெஸ்டில் (Reader's Digest) நிறுவப்பட்டது. கணினியில் உள்ளீடு செய்வதற்காகத் தட்டச்சு செய்யப்பட்ட விற்பனை அறிக்கைகளைப் பஞ்ச் கார்டுகளாக மாற்ற இந்த உபகரணம் பயன்படுத்தப்பட்டது.

3) முதல் தலைமுறை எழுத்துணரி

ஆயிரத்து தொள்ளாயிரத்து அறுபது முதல் ஆயிரத்து தொள்ளாயிரத்து அறுபதைந்து வரையிலான காலகட்டத்தில் தோன்றிய வணிக எழுத்துணரி ஒழுங்குமுறைகளை

எழுத்துணரியின் முதல் தலைமுறை என்று அழைக்கலாம். இந்தத் தலைமுறை எழுத்துணர் இயந்திரங்கள் முக்கியமாக வாசிக்கப்பட்ட கட்டுப்படுத்தப்பட்ட எழுத்து வடிவங்களால் வகைப்படுத்தப்பட்டன. சின்னங்கள் இயந்திர வாசிப்புக்காக பிரத்யேகமாக வடிவமைக்கப்பட்டன; முதல் வகைகள் மிகவும் இயல்பாகத் தெரியவில்லை. காலப்போக்கில் பத்து வெவ்வேறு எழுத்துருக்களைப் படிக்கக்கூடும் பல எழுத்துரு இயந்திரங்கள் தோன்றத் தொடங்கின. எழுத்துருக்களின் எண்ணிக்கைப் பயன்படுத்தப்பட்ட அமைப்பொழுங்கு உணர்தல் முறை, வார்ப்புரு பொருத்தம் ஆகியவற்றால் எல்லைப்படுத்தப்பட்டிருந்தது; இது எழுத்துருப் படத்தை ஒவ்வொரு எழுத்துருவின் ஒவ்வொரு எழுத்துக்கும் முன்மாதிரி படங்களின் நூலகத்துடன் ஒப்பிட்டது.

4) இரண்டாம் தலைமுறை எழுத்துணரி

இரண்டாம் தலைமுறையின் வாசிப்பு இயந்திரங்கள் ஆயிரத்து தொள்ளாயிரத்து அறுபதுகளின் நடுப்பகுதியிலும் ஆயிரத்து தொள்ளாயிரத்து எழுபதுகளின் முற்பகுதியிலும் தோன்றின. இந்த ஒழுங்குமுறைகளால் வழக்கமான இயந்திர அச்சிடப்பட்ட எழுத்துக்களை அடையாளம் காண முடிந்தது, மேலும் கையால் அச்சிடப்பட்ட எழுத்துக்குறி உணர்தல் திறன்களையும் கொண்டிருந்தன. கையால் அச்சிடப்பட்ட எழுத்துக்கள் கருதப்பட்டபோது, எழுத்துக்குறி தொகுப்பு எண்கள் மற்றும் ஒரு சில எழுத்துக்கள் மற்றும் சின்னங்களுடன் கட்டுப்படுத்தப்பட்டது.

இந்த வகையான முதல் மற்றும் பிரபலமான அமைப்பு ஐபிஎம் 1287 ஆகும்; இது ஆயிரத்து தொள்ளாயிரத்து அறுபத்து ஐந்தாம் ஆண்டில் நியூயார்க்கில் நடந்த உலக கண்காட்சியில் காட்சிக்கு வைக்கப்பட்டது. மேலும், இந்த காலகட்டத்தில் தோஷிபா அஞ்சல் குறியீடு எண்களுக்கான முதல் தானியங்கி கடிதம் வரிசைப்படுத்தும் இயந்திரத்தை

உருவாக்கியது மற்றும் அதிக செயல்திறன் மற்றும் குறைந்த செலவில் ஹிட்டாச்சி முதல் எழுத்துணரி இயந்திரத்தை உருவாக்கியது.

இந்த காலகட்டத்தில் தரநிலைப்படுத்தல் பகுதியில் குறிப்பிடத்தக்க பணிகள் மேற்கொள்ளப்பட்டன. ஆயிரத்து தொள்ளாயிரத்து அறுபத்து ஆறாம் ஆண்டில், எழுத்துணரி தேவைகள் பற்றிய முழுமையான ஆய்வு முடிக்கப்பட்டது மற்றும் ஒரு தரமான அமெரிக்க எழுத்துணரி எழுத்துக்குறித் தொகுப்பு (American standard OCR character set) எழுத்துணரி-A வரையறுக்கப்பட்டது. இந்த எழுத்துரு மனிதர்களுக்கு இன்னும் படிக்கக்கூடியதாக இருந்தாலும் மிகவும் பகட்டானது மற்றும் ஒளிவழி உணர்தலை எளிதாக்க வடிவமைக்கப்பட்டது. ஒரு ஐரோப்பிய எழுத்துருவும் எழுத்துணரி-B வடிவமைக்கப்பட்டது; இது அமெரிக்கத் தரத்தை விட இயற்கையான எழுத்துருக்களைக் கொண்டிருந்தது. இரண்டு எழுத்துருக்களையும் ஒரே தரத்தில் இணைக்க சில முயற்சிகள் மேற்கொள்ளப்பட்டன; ஆனால் அதற்குப் பதிலாக இரண்டு தரங்களையும் படிக்கக்கூடிய இயந்திரங்கள் தோன்றின.

5) மூன்றாம் தலைமுறை எழுத்துணரி

ஆயிரத்து தொள்ளாயிரத்து எழுபதுகளின் நடுப்பகுதியில் தோன்றிய மூன்றாம் தலைமுறை எழுத்துணரி அமைப்புகளுக்குச் சவாலாக அமைந்தது மோசமான தரம் மற்றும் பெரிய அச்சிடப்பட்ட மற்றும் கையால் எழுதப்பட்ட எழுத்துத் தொகுப்புகளின் ஆவணங்கள் ஆகும். குறைந்த செலவு மற்றும் அதிக செயல்திறன் ஆகியவை முக்கியமான நோக்கங்களாக இருந்தன; அவை வன்பொருள் தொழில்நுட்பத்தின் வியத்தகு முன்னேற்றங்களால் உதவப்பட்டன.

மிகவும் அதிநவீன எழுத்துணரி-இயந்திரங்கள் சந்தையில் தோன்றத் தொடங்கினாலும் எளிய எழுத்துணரிக் கருவிகள் இன்னும் மிகவும் பயனுள்ளதாக

இருந்தன. தனிநபர் கணினிகள் மற்றும் லேசர் அச்சுப்பொறிகள் உரை உற்பத்தியில் ஆதிக்கம் செலுத்தத் தொடங்குவதற்கு முந்தைய காலகட்டத்தில், தட்டச்சு செய்வது எழுத்துணரிக்கு ஒரு சிறப்பு இடமாகும். ஒரே மாதிரியான அச்சு இடைவெளி மற்றும் சிறிய எண்ணிக்கையிலான எழுத்துருக்கள் எளிதாக வடிவமைக்கப்பட்ட எழுத்துணரிக் கருவிகளை மிகவும் பயனுள்ளதாக ஆக்கியது. சாதாரணத் தட்டச்சுப்பொறிகளில் கரடுமுரடான வரைவுகளை உருவாக்கலாம் மற்றும் இறுதி திருத்தம் (எடிட்டிங்) செய்ய எழுத்துணரிக் கருவி மூலம் கணினியில் செலுத்தலாம். இந்த வழியில் இந்த நேரத்தில் விலையுயர்ந்த வளமாக இருந்தன வேர்ட் செயலிகள் பலவற்றை ஆதரிக்கக்கூடும்; மேலும் உபகரணங்களுக்கான செலவுகளைக் குறைக்க இயலும்.

6) எழுத்துணரி இன்று

இருப்பினும், ஆயிரத்து தொள்ளாயிரத்து ஐம்பதுகளில் எழுத்துணரி இயந்திரங்கள் வணிக ரீதியாகக் கிடைத்தாலும், ஆயிரத்து தொள்ளாயிரத்து எண்பத்தாறு வரை உலகளவில் சில ஆயிரம் ஒழுங்குமுறைகள் மட்டுமே விற்கப்பட்டன. இதற்கு முக்கிய காரணம் ஒழுங்குமுறைகளின் விலை. இருப்பினும், வன்பொருள் மலிவாகி வருவதாலும், எழுத்துணரி ஒழுங்குமுறைகள் மென்பொருள் தொகுப்புகளாகக் கிடைக்கத் தொடங்கியதாலும், விற்பனை கணிசமாக அதிகரித்தது. இன்று ஒரு சில ஆயிரம் என்பது ஒவ்வொரு வாரமும் விற்கப்படும் ஒழுங்குமுறைகளின் எண்ணிக்கையாகும், மேலும் ஓம்னி எழுத்துரு எழுத்துணரியின் விலை கடந்த 6 ஆண்டுகளாக ஒவ்வொரு ஆண்டும் பத்து என்ற காரணியுடன் குறைந்துள்ளது.

காலம்	**முன்னேற்றம்**

1870	முதல் முயற்சிகள்
1940	எழுத்துணரியின் தற்கால பதிப்பு
1950	முதல் எழுத்துணரி இயந்திரத்தின் தோற்றம்
1960 - 1965	முதல் தலைமுறை எழுத்துணரி
1965 - 1975	இரண்டாம் தலைமுறை எழுத்துணரி
1975 - 1985	மூன்றாம் தலைமுறை எழுத்துணரி
1986 ->	மக்களுக்கான எழுத்துணரி

ஆரம்பகால ஒளிமூலம் எழுத்துணரி தந்தி சம்பந்தப்பட்ட தொழில்நுட்பங்கள் மற்றும் பார்வையற்றோருக்கான வாசிப்புச் சாதனங்களை உருவாக்குதல் ஆகியவற்றைக் கண்டறியலாம் (Schantz, 1982). ஆயிரத்து தொள்ளாயிரத்துப் பதிநான்காம் ஆண்டில், இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg) ஒரு இயந்திரத்தை உருவாக்கி, எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றினார் (Dhavale, 2017). ஒரே நேரத்தில், எட்மண்ட் ஃபோர்னியர் டி ஆல்பே (Edmund Fournier d'Albe) ஒரு கையடக்க ஸ்கேனரான ஆப்டோஃபோனை (Optophone) உருவாக்கினார், இது அச்சிடப்பட்ட பக்கத்தின் குறுக்கே நகரும்போது, குறிப்பிட்ட எழுத்துக்கள் அல்லது எழுத்துக்களுக்கு ஒத்த டோன்களை உருவாக்கியது (d'Albe, 1914).

ஆயிரத்து தொள்ளாயிரத்து இருபதுகளின் பிற்பகுதியிலும் ஆயிரத்து தொள்ளாயிரத்து முப்பதுகளிலும் இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg) ஒரு ஆப்டிகல் குறியீடு புரிந்துகொள்ளும் முறையைப் பயன்படுத்தி மைக்ரோஃபில்ம் காப்பகங்களைத் தேடுவதற்காக "புள்ளிவிவர இயந்திரம்" என்று அழைக்கப்பட்ட ஒன்றை உருவாக்கினார். தொள்ளாயிரத்து முப்பது ஒன்றாம் ஆண்டில் அவரது கண்டுபிடிப்புக்காக

அமெரிக்காவின் காப்புரிமை எண் 1,838,389 வழங்கப்பட்டது. காப்புரிமையை ஐ.பி.எம்.-ஆல் பெறப்பட்டது.

பார்வையற்ற மற்றும் பார்வைக் குறையுள்ள பயனர்கள்

ஆயிரத்து தொள்ளாயிரத்து எழுபத்து நாலாம் ஆண்டில், ரே குர்ஸ்வீல் (Ray Kurzweil) குர்ஸ்வீல் கம்ப்யூட்டர் தயாரிப்புகள், இன்க் (Kurzweil Computer Products, Inc.) என்ற நிறுவனத்தைத் தொடங்கினார். மற்றும் ஓம்னி-எழுத்துரு ஒளிமூலம் எழுத்து உணரியின் (omni-font OCR) தொடர்ச்சியான உருவாக்கத்தைத் தொடர்ந்தார், இது எந்தவொரு எழுத்துருவிலும் அச்சிடப்பட்ட உரையை உணர இயலும் (குர்ஸ்வீல் பெரும்பாலும் ஓம்னி-எழுத்துரு ஒளிமூலம் எழுத்து உணரியைக் கண்டுபிடித்த பெருமைக்கு உள்ளாகிறார்; அது ஆயிரத்து தொள்ளாயிரத்து அறுபதுகளின் பிற்பகுதியிலும் ஆயிரத்து தொள்ளாயிரத்து எழுபத்துகளில் பிற்பகுதியிலும் கம்ப்யூஸ்கான் (CompuScan) உள்ளிட்ட நிறுவனங்களின் பயன்பாட்டில் இருந்தது). பார்வையற்றோருக்கு ஒரு வாசிப்பு இயந்திரத்தை உருவாக்குவதே இந்தத் தொழில்நுட்பத்தின் சிறந்த பயன்பாடாகும் என்று குர்ஸ்வீல் முடிவு செய்தார், இது பார்வையற்றவர்களுக்குக் கணினி வாசிக்கும் உரையைச் சத்தமாக வைத்திருக்க அனுமதிக்கும் ("The History of OCR"). இந்தச் சாதனத்திற்கு இரண்டு செயல்படுத்தும் தொழில்நுட்பங்களின் கண்டுபிடிப்பு தேவை - சிசிடி பிளாட்பெட் ஸ்கேனர் மற்றும் உரையிலிருந்து பேச்சு உருவாக்கம். ஜனவரி 13, 1976 அன்று, குர்ஸ்வீல் மற்றும் பார்வையற்றோரின் தேசியக் கூட்டமைப்பின் தலைவர்கள் தலைமையில் பரவலாக அறிவிக்கப்பட்ட செய்தி மாநாட்டின் போது வெற்றிகரமாக முடிக்கப்பட்ட தயாரிப்பு வெளியிடப்பட்டது. ஆயிரத்து தொள்ளாயிரத்து எழுபத்து எட்டாம் ஆண்டில், குர்ஸ்வீல் கணினித் தயாரிப்புகள் ஒளிவழி எழுத்து உணர்தல் கணினி நிரல் (optical

character recognition computer program) வணிகப் பதிப்பை விற்பனை செய்யத் தொடங்கின. லெக்சிஸ்நெக்ஸிஸ் (LexisNexis) முதல் வாடிக்கையாளர்களில் ஒன்றாக இருந்தது; மேலும் சட்ட ஆவணங்கள் மற்றும் செய்தி ஆவணங்களை அதன் புதிய நிகழ்நிலை (ஆன்லைன்) தரவுத்தளங்களில் பதிவேற்றுவதற்கான திட்டத்தை வாங்கினார். இரண்டு ஆண்டுகளுக்குப் பிறகு, குர்ஸ்வீல் தனது நிறுவனத்தை ஜெராக்ஸுக்கு விற்றார்; ஜெராக்ஸ் காகிதத்திலிருந்து கணினி உரை மாற்றத்தை மேலும் வணிகமயமாக்குவதில் ஆர்வம் கொண்டிருந்தது. ஜெராக்ஸ் இறுதியில் அதை ஸ்கேன்சாஃப்ட் (Scansoft) என்று அழைத்தது; இது நுவான்ஸ் கம்யூனிகேஷனுடன் (Nuance Communications) இணைந்தது.

இரண்டாயிரங்களில் ஒளிமூலம் எழுத்துணரி நிகழ்நிலையில் (ஆன்லைனில்) ஒரு சேவையாக (WebOCR), கிளவுட் கம்ப்யூட்டிங்/மேகக் கணிப்பு சூழலில் மற்றும் திறன்பேசியில் (ஸ்மார்ட்போன்/smartphone) வெளிநாட்டு மொழிக் குறியீகளின் (foreign-language signs) நிகழ்நேர மொழிபெயர்ப்பு போன்ற கைபேசிப் (மொபைல்/Mobile) பயன்பாடுகளில் கிடைக்கச் செய்யப்பட்டது. திறன்பேசிகள் (ஸ்மார்ட் போன்கள்) மற்றும் திறன் கண்ணாடிகள் (ஸ்மார்ட் கிளாஸ்கள்/Smart glasses) இவற்றின் வருகையுடன், கருவியின் புகைப்படக்கருவியைப் (camera/கேமரா) பயன்படுத்திப் பெறப்பட்ட உரையைப் பிரித்தெடுக்கும் இணையம் எழுத்துணரியைப் பயன்படுத்தலாம். இயக்க முறைமையில் (operating system) கட்டமைக்கப்பட்ட எழுத்துணரி செயல்பாடு இல்லாத இந்தக் கருவிகள் பொதுவாக ஒளிமூலம் எழுத்துணரி எபிஐ-ஐப் (OCR API) பயன்படுத்தி கருவியால் பெறப்பட்ட மற்றும் வழங்கப்பட்ட படக் கோப்பிலிருந்து உரையைப் பிரித்தெடுக்கும். அசல் படத்தில் கண்டறியப்பட்ட உரையின் இருப்பிடம் பற்றிய

தகவலுடன், மேலும் செயலாக்கத்திற்காக (உரையிலிருந்து பேச்சு போன்றவை) அல்லது காட்சிக்குக் கருவிப் பயன்பாட்டிற்குத் எழுத்துணரி எபிஐ (OCR API) திரும்பப் பெறுகிறது.

இலத்தீன், சிரிலிக், அரபு, ஹீப்ரு, இந்திக், பெங்காலி (பங்களா), தேவநாகரி, தமிழ், சீன, ஜப்பானிய மற்றும் கொரிய எழுத்துக்கள் உள்ளிட்ட பொதுவான வணிக அமைப்புகளுக்குப் பல்வேறு வணிக மற்றும் திறந்த மூல ஒளிமூலம் எழுத்துணரி ஒழுங்குமுறைகள் கிடைக்கின்றன.

நுட்பங்கள் (Techniques)

1) முன் செயலாக்கம் (Pre-processing)

ஒளிமூலம் எழுத்துணரி (ஓசிஆர்/OCR) மென்பொருள் பெரும்பாலும் வெற்றிகரமான உணர்தலுக்கான வாய்ப்புகளை மேம்படுத்த படங்களை "முன் செயலாக்கம் செய்கிறது" ("pre-processes"). நுட்பங்கள் பின்வருமாறு (Optical Character Recognition (OCR) – How it works". Nicomsoft.com):

- டி-ஸ்கேவ் (De-skew) - வருடும் (ஸ்கேன்) போது ஆவணம் சரியான திசையில் சாய்ந்து கொள்ள வேண்டியிருக்கும்.
- டெஸ்பெகிள் (Despeckle) - நேர்மறை மற்றும் எதிர்மறைப் புள்ளிகள், மென்மையான விளிம்புகளை அகற்றும்.
- இருமையாக்கம் (Binarisation) - ஒரு படத்தை வண்ணம் அல்லது கிரேஸ்கேலில் (greyscale) இருந்து கருப்பு மற்றும் வெள்ளை நிறமாக மாற்றவும் (இரண்டு வண்ணங்கள் இருப்பதால் "பைனரி படம்" ("binary image") என்று அழைக்கப்படுகிறது)). உரையை (அல்லது வேறு ஏதேனும் விரும்பிய படக் கூறுகளை) பின்னணியில் இருந்து பிரிப்பதற்கான எளிய வழியாக இருமையாக்கப்

(பைனரைசேஷன்) பணி செய்யப்படுகிறது (Sezgin & Sankur 2004). பெரும்பாலான வணிக உணர்தல் வழிமுறைகள் இருமைப் (பைனரி) படங்களில் மட்டுமே செயல்படுவதால் இருமையாக்கப் (binarisation) பணி அவசியம். ஏனெனில் அவ்வாறு செய்வது எளிது என்பது நிரூபிக்கப்பட்டுள்ளது (Gupta, Maya et al 2007). கூடுதலாக, இருமையாக்கம் (binarisation) படியின்/நடவடிக்கையின் செயல்திறன் ஒரு குறிப்பிடத்தக்க அளவிற்கு எழுத்துப் புரிதல் கட்டத்தின் தரத்தை அதிகாரம்செய்கிறது; மற்றும் ஒரு குறிப்பிட்ட உள்ளீட்டு பட வகைக்குப் பயன்படுத்தப்படும் இருமையாக்கத்தை (binarisation) தேர்ந்தெடுப்பதில் கவனமாக முடிவுகள் எடுக்கப்படுகின்றன; இருமை (binary) முடிவைப் பெறுவதற்குப் பயன்படுத்தப்படும் இருமையாக்க (பைனரைசேஷன்) முறையின் தரம் உள்ளீட்டுப் படத்தின் வகையைப் பொறுத்தது (ஸ்கேன் செய்யப்பட்ட ஆவணம், காட்சி உரை படம், வரலாற்று சீரழிந்த ஆவணம் போன்றவை) (Trier & Jain 1995; Milyaev et al 2013).

- வரி நீக்கம் (Line removal) - கிளிஃப் அல்லாத பெட்டிகளையும் கோடுகளையும் சுத்தம் செய்கிறது
- தளவமைப்பு பகுப்பாய்வு அல்லது "ஸோனிங்" (Layout analysis or "zoning") - நெடுவரிசைகள், பத்திகள், தலைப்புகள் போன்றவற்றைத் தனித்துவமான தொகுதிகளாக அடையாளம் காட்டுகிறது. பல நெடுவரிசை தளவமைப்புகள் மற்றும் அட்டவணைகளில் குறிப்பாக முக்கியமானது.
- வரி மற்றும் சொல் கண்டறிதல் (Line and word detection) - சொல் மற்றும் எழுத்து வடிவங்களுக்கான அடிப்படைகளை நிறுவுகிறது, தேவைப்பட்டால் சொற்களைப் பிரிக்கிறது.

- எழுத்துவடிவத்தைப் புரிதல் (Script recognition) - பன்மொழி ஆவணங்களில், எழுத்துவடிவம் சொற்களின் மட்டத்தில் மாறக்கூடும், எனவே, குறிப்பிட்ட எழுத்துவடிவத்தைக் கையாளச் சரியான எழுத்துணரியைப் (OCR) பயன்படுத்துவதற்கு முன்பு, எழுத்துவடிவத்தை அடையாளம் காண்பது அவசியம்.
- எழுத்து தனிமைப்படுத்தல் அல்லது "கூறுபடுத்தல்" (Character isolation or "segmentation") - ஒவ்வொரு எழுத்துக்குறி ஒளிமூலம் எழுத்துணரிக்கு (per-character OCR), படக் கலைப்பொருட்கள் (image artifacts) காரணமாக இணைக்கப்பட்டுள்ள பல எழுத்துக்கள் பிரிக்கப்பட வேண்டும்; செயற்கைப்பொருள் (artifacts) காரணமாகப் பல துண்டுகளாக உடைக்கப்பட்ட ஒற்றை எழுத்துக்கள் இணைக்கப்பட வேண்டும்.
- தோற்ற விகிதம் மற்றும் அளவை இயல்பாக்குதல் நிலையான-இசைமை எழுத்துருக்களின் (fixed-pitch fonts) கூறாக்கம் ஒப்பீட்டளவில் வெறுமனே படத்தை ஒரு சீரான கட்டத்துடன் (uniform grid) செங்குத்து கட்டக் கோடுகள் எப்போதாவது கருப்புப் பகுதிகளை இடைவெட்டுகின்றன என்பதன் அடிப்படையில் இணைப்பதன் மூலம் நிறைவேற்றப்படுகிறது. விகிதாசார எழுத்துருக்களுக்கு, அதிநவீன நுட்பங்கள் தேவைப்படுகின்றன, ஏனென்றால் எழுத்துக்களுக்கு இடையில் உள்ள இடைவெளி சில நேரங்களில் சொற்களுக்கு இடையில் இருப்பதை விட அதிகமாக இருக்கலாம், மேலும் செங்குத்து கோடுகள் ஒன்றுக்கு மேற்பட்ட எழுத்துக்களை இடைவெட்டக்கூடும் (Ray Smith 2007).

2) உரை புரிதல் (Text recognition)

மைய ஒளிமூலம் எழுத்துப் புரிவான் (கோர் ஓ.சி.ஆர்.) வழிமுறையின் இரண்டு அடிப்படை வகைகள் உள்ளன, அவை தேர்வுக்குரிய எழுத்துக்களின் தரவரிசைப் பட்டியலை உருவாக்கக்கூடும் ("OCR Introduction". Dataid.com.)

மேட்ரிக்ஸ் பொருத்தம் (Matrix matching) என்பது ஒரு படத்தை பிக்சல்-பை-பிக்சல் அடிப்படையில் சேமிக்கப்பட்ட கிளிஃபுடன் ஒப்பிடுவதை உள்ளடக்குகிறது; இது "அமைப்பொழுங்குப் பொருத்தம்" ("pattern matching"), "அமைப்பொழுங்குப் புரிதல்" ("pattern recognition") அல்லது "படத் தொடர்பு", ("image correlation") என்றும் அழைக்கப்படுகிறது. இது உள்ளீட்டுக் கிளிஃபைப் படத்தின் பிற பகுதிகளிலிருந்து சரியாகத் தனிமைப்படுத்தப்படுவதையும், சேமிக்கப்பட்ட கிளிஃப் ஒத்த எழுத்துருவிலும் அதே அளவிலும் இருப்பதையும் நம்பியுள்ளது. இந்த நுட்பம் தட்டச்சு செய்யப்பட்ட உரையுடன் சிறப்பாகச் செயல்படுகிறது, மேலும் புதிய எழுத்துருக்களை எதிர்கொள்ளும்போது நன்றாக வேலை செய்யாது. ஆரம்பகால இயற்பியல் ஒளிச்சேர்க்கை அடிப்படையிலான எழுத்துணரி (physical photocell-based OCR) நேரடியாகச் செயல்படுத்தப்பட்ட நுட்பமாகும்.

பண்புக்கூறு பிரித்தெடுத்தல் கோடுகள் (lines), மூடிய கண்ணிகள்/வளையங்கள் (closed loops), வரி திசை (line direction) மற்றும் வரி குறுக்குவெட்டுகள் (line intersections) போன்ற "பண்புக்கூறுகளாக" கிளிஃப்களைச் சிதைக்கிறது. பிரித்தெடுக்கும் பண்புக்கூறுகள் உருப்படுத்தத்தின் பரிமாணத்தைக் குறைக்கிறது மற்றும் புரிதல் செயல்முறையைக் கணினி/கணக்கீட்டு அடிப்படையில் திறம்படச் செய்கிறது. இந்தப் பண்புக்கூறுகள் ஒரு எழுத்தின் சுருக்கத் திசையன் போன்ற (abstract vector-like) உருப்படுத்தத்துடன் ஒப்பிடப்படுகின்றன, இது ஒன்று அல்லது அதற்கு மேற்பட்ட கிளிஃப்

மூலமுன்மாதிரிகளாகக் (glyph prototypes) குறையக்கூடும். கணினிப் பார்வையில் பண்புக்கூறைக் கண்டறிவதற்கான பொதுவான நுட்பங்கள் இந்த வகை ஒளி எழுத்துணரிக்குப் (ஓசிஆர்/OCR) பொருந்தும், இது பொதுவாக "அறிவார்ந்த" கையெழுத்துப் புரிதல் ("intelligent" handwriting recognition) மற்றும் உண்மையில் நவீன எழுத்துணரி (ஓசிஆர்/OCR) மென்பொருளில் காணப்படுகிறது. கே-அருகிலுள்ள அண்டைகள் வழிமுறை (k-nearest neighbors algorithm) போன்ற அருகிலுள்ள அண்டை வகைப்படுத்திகள் (Nearest neighbour classifiers) படப் பண்புக்கூறுகளைச் சேமிக்கப்பட்ட கிளிஃப் பண்புக்கூறுகளுடன் ஒப்பிட்டு அருகிலுள்ள பொருத்ததைத் (nearest match) தேர்வுசெய்யப் பயன்படுத்தப்படுகின்றன.

கியூனிஃபார்ம் (Cuneiform) மற்றும் டெசராக்ட் (Tesseract) போன்ற மென்பொருள்கள் எழுத்துப் புரிதலுக்கு (character recognition) இரண்டு-பாஸ் அணுகுமுறையைப் பயன்படுத்துகின்றன. இரண்டாவது பாஸ் (two-pass approach) "தகவமைப்பு புரிதல்" ("adaptive recognition") என்று அழைக்கப்படுகிறது மற்றும் இரண்டாவது பாஸில் மீதமுள்ள எழுத்துக்களைச் சிறப்பாக அடையாளம் காண முதல் பாஸில் அதிக நம்பிக்கையுடன் புரியப்பட்ட எழுத்து வடிவங்களைப் பயன்படுத்துகிறது. எழுத்துரு சிதைந்த அசாதாரண எழுத்துருக்கள் அல்லது குறைந்த தரமான ஸ்கேன்களுக்கு (எ.கா. மங்கலான அல்லது மங்கிப்போன) இது சாதகமானது (Ray Smith 2007).

நவீன ஒளிமூலம் எழுத்துப் புரிதல் (OCR) மென்பொருளானது ஓசிஆர்ஓபஸ் OCRopus அல்லது தெஸ்ஸெராக்ட் (Tesseract) போன்ற நரம்பியல் வலையமைப்புகளைப் (நெட்வொர்க்குகளைப்) பயன்படுத்துகிறது, அவை ஒற்றை எழுத்துக்களில் கவனம் செலுத்துவதற்குப் பதிலாக உரையின் முழு வரிகளையும் அடையாளம் காணப் பயிற்சி பெற்றன.

ஓசிஆர்/OCR முடிவை தரப்படுத்தப்பட்ட ஆல்டோ (ALTO) வடிவத்தில் சேமிக்க முடியும், இது யுனைடெட் ஸ்டேட்ஸ் லைப்ரரி ஆஃப் காங்கிரஸால் (United States Library of Congress) பராமரிக்கப்படும் ஒரு பிரத்யேக எக்ஸ்எம்எல் (XML) திட்டமாகும். பிற பொதுவான வடிவங்களில் hOCR மற்றும் PAGE XML ஆகியவை அடங்கும்.

ஒளிவழி எழுத்துப் உணர்தல் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) மென்பொருளின் பட்டியலுக்கு ஒளிவழி எழுத்துப் உணர்தல் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) மென்பொருளின் ஒப்பீடு பார்க்கவும்.

3) பின் செயலாக்கம் (Post-processing)

வெளியீட்டை ஒரு அகராதி (ஒரு ஆவணத்தில் நிகழ அனுமதிக்கப்பட்ட சொற்களின் பட்டியல்) மூலம் கட்டுப்படுத்தினால் ஒளிமூலம் எழுத்துப் புரிவான் துல்லியத்தை அதிகரிக்க இயலும் ("Optical Character Recognition (OCR) – How it works". Nicomsoft.com.). எடுத்துக்காட்டாக, இது ஆங்கில மொழியில் உள்ள அனைத்து சொற்களும் அல்லது ஒரு குறிப்பிட்ட புலத்திற்கான தொழில்நுட்ப அகராதியாக இருக்கலாம். இயற்பெயர்களைப் போல, அகராதியில் இல்லாத சொற்கள் ஆவணத்தில் இருந்தால் இந்த நுட்பம் சிக்கலாக இருக்கும். மேம்பட்ட துல்லியத்திற்காக, எழுத்து கூறாக்க நடத்தையை ஊகுவிக்க டெசராக்ட் (Tesseract) அதன் அகராதியைப் பயன்படுத்துகிறது (Ray Smith 2007).

வெளியீட்டு ஒழுக்கு ஒரு எளிய உரை ஒழுக்கு அல்லது எழுத்துகளின் கோப்பாக இருக்கலாம், ஆனால் மிகவும் அதிநவீன ஒளிவழி எழுத்துணரி அமைப்புகள் பக்கத்தின் அசல் தளவமைப்பைப் பாதுகாத்து உருவாக்கலாம்; எடுத்துக்காட்டாக, பக்கத்தின் அசல்

படம் மற்றும் தேடக்கூடிய உரை உருப்படுத்தம் ஆகிய இரண்டையும் உள்ளடக்கிய ஒரு அடையாளப்படுத்தப்பட்ட பிடிஎஃ (PDF).

"அருகிலுள்ள அண்டைப் பகுப்பாய்வு" ("Near-neighbor analysis") சில சொற்கள் பெரும்பாலும் ஒன்றாகக் காணப்படுவதைக் குறிப்பிடுவதன் மூலம் பிழைகளைச் சரிசெய்ய இணைநிகழ்வு அதிர்வெண்களைப் (co-occurrence frequencies) பயன்படுத்தலாம் ("How does OCR document scanning work?"). எடுத்துக்காட்டாக, "Washington, D.C." "Washington DOC"-ஐ விட ஆங்கிலத்தில் மிகவும் பொதுவானது.

வருடல்/ஸ்கேன் செய்யப்படும் மொழியின் இலக்கணத்தைப் பற்றிய அறிவு ஒரு சொல் வினைச்சொல்லாகவோ அல்லது பெயர்ச்சொல்லாகவோ இருக்க முடியுமா என்பதை தீர்மானிக்க உதவும், எடுத்துக்காட்டாக, அதிக துல்லியத்தை அனுமதிக்கிறது.

எழுத்துணரி எபிஐ-இன் (OCR API) முடிவுகளை மேலும் மேம்படுத்த ஒளிமூலம் எழுத்துப் புரிவான் பிந்தைய செயலாக்கத்திலும் லெவன்ஸ்டீன் தொலைநிலை வழிமுறை (Levenshtein Distance algorithm) பயன்படுத்தப்பட்டுள்ளது ("How to optimize results from Community").

4) பயன்பாடு சார்ந்த மேம்படுத்தல்கள் (Application-specific optimizations)

சமீபத்திய ஆண்டுகளில், முக்கிய ஒளிமூலம் எழுத்துணரி தொழில்நுட்ப வழங்குநர்கள் குறிப்பிட்ட வகை உள்ளீடுகளை மிகவும் திறமையாகக் கையாள ஒளிமூலம் எழுத்துப் புரிதல் ஒழுங்குமுறைகளை மாற்றத் தொடங்கினர். பயன்பாட்டு-குறிப்பிட்ட அகராதிக்கு (application-specific lexicon) அப்பால், வணிக விதிகள், நிலையான வெளிப்பாடு, அல்லது வண்ணப் படங்களில் உள்ள வளமான தகவல்களைக் கணக்கில் எடுத்துக்கொள்வதன் மூலம் சிறந்த செயல்திறனைப் பெறலாம். இந்த மூலோபாயம் "பயன்பாடு சார்ந்த ஒளிமூலம் எழுத்துணரி" ("Application-Oriented OCR") அல்லது

"தனிப்பயனாக்கப்பட்ட ஒளிமூலம் எழுத்துணரி" ("Customized OCR") என்று அழைக்கப்படுகிறது; மேலும் இது உரிமத் தகடுகள், விலைப்பட்டியல், திரை தனிக்காட்சிகள் (ஸ்கிரீன் ஷாட்கள்), அடையாள அட்டைகள், ஓட்டுநர் உரிமங்கள் மற்றும் தானியங்கியூர்தி (ஆட்டோமொபைல்) உற்பத்தி ஆகியவற்றில் ஒளிமூலம் எழுத்துணரி பயன்படுத்தப்படுகிறது.

நியூயார்க் டைம்ஸ் எழுத்துணைரி (ஓசிஆர்/OCR) தொழில்நுட்பத்தை அவர்கள் வைத்திருக்கும் தனியுரிமக் கருவியாக ஆவண உதவியாளராக (Document Helper) மாற்றியமைத்துள்ளது, இது அவர்களின் ஊடாடும் செய்திக் குழுவை மதிப்பாய்வு செய்ய வேண்டிய ஆவணங்களின் செயலாக்கத்தை துரிதப்படுத்த உதவுகிறது. நிருபர்கள் உள்ளடக்கங்களை மறுஆய்வு செய்வதற்கான தயாரிப்பில் ஒரு மணி நேரத்திற்கு 5,400 பக்கங்கள் வீதம் செயலாக்க இது உதவுகிறது என்பதை அவர்கள் குறிப்பிடுகிறார்கள் (Fehr 2019)

வகைகள்

கீழ் வருவன ஒளிவழி எழுத்துணரியின் (ஆப்டிகல் கேரக்டர் ரெக்னிஷன் (OCR)) வகைகளாகப் பட்டியலிடப்பட்டுள்ளன.

- ஒளிவழி எழுத்து உணர்தல் (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் Optical character recognition (OCR)) - தட்டச்சு செய்யப்பட்ட உரை குறிவைக்கிறது; ஒரே நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்து என்ற முறையில்.
- ஒளிவழி சொல் உணர்தல் (Optical word recognition) - தட்டச்சு செய்யப்பட்ட உரையைக் குறிவைக்கிறது; ஒரே நேரத்தில் ஒரு சொல் என்ற முறையில் (இடத்தை

சொல் பிரிப்பியாகப் பயன்படுத்தும் மொழிகளுக்கு). (பொதுவாக "எழுதுணரி (OCR)" என்று அழைக்கப்படுகிறது.)

- நுண்ணறிவு எழுத்துப் புரிதல் (Intelligent character recognition (ICR)) - கையால் எழுதப்பட்ட அச்சுஎழுத்து (handwritten printscript) அல்லது இணைவெழுத்து (கர்சீவ்) உரையைக் குறிவைக்கிறது (cursive text); ஒரு நேரத்தில் ஒரு கிளிஃப் (glyph) அல்லது எழுத்து என்ற முறையில்; பொதுவாக இயந்திரக் கற்றல் இதில் அடங்கும்.
- நுண்ணறிவுச் சொல் புரிதல் (Intelligent Word recognition (IWR)) - கையால் எழுதப்பட்ட அச்சுஎழுத்து அல்லது இணைவெழுத்து (கர்சீவ்) உரையைக் குறிவைக்கிறது; ஒரு நேரத்தில் ஒரு சொல் என்ற முறையில். இணைவெழுத்து (கர்சீவ்) எழுத்துவடிவில் கிளிஃப்கள் (glyphs) கூறிடப்படாத மொழிகளுக்கு இது மிகவும் பயனுள்ளதாக இருக்கும்.

எழுத்துணரி (ஓசிஆர்) பொதுவாக ஒரு "நிகழாநிலை" ("ஆஃப்லைன்" செயல்முறையாகும், இது ஒரு நிலையான ஆவணத்தை பகுப்பாய்வு செய்கிறது. ஆன்லைன் ஒளிவழி எழுத்துப் உணர்தல் எபிஐ (OCR API) சேவையை வழங்கும் கிளவுட் அடிப்படையிலான சேவைகள் உள்ளன. கையெழுத்து இயக்கப் பகுப்பாய்வு (Handwriting movement analysis), கையெழுத்துப் புரிதலுக்கான (handwriting recognition) உள்ளீடாகப் பயன்படுத்தப்படலாம் (Tappert et al 1990). கிளிஃப்கள் மற்றும் சொற்களின் வடிவங்களைப் பயன்படுத்துவதற்குப் பதிலாக, இந்த நுட்பம், கூறுகள் வரையப்பட்ட வரிசை, திசை, மற்றும் பேனாவைக் கீழே வைத்து தூக்கும் அமைபொழுங்கு போன்ற

இயக்கங்களை ஈட்ட இயலும். இந்தக் கூடுதல் தகவல் இறுதிக்கு-இறுதி செயல்முறையை மிகவும் துல்லியமாக்குகிறது. இந்தத் தொழில்நுட்பம் "ஆன்-லைன் எழுத்துணர்தல்" ("on-line character recognition"), "இயங்கு எழுத்து உணர்தல்" ("dynamic character recognition"), "நிகழ்நேர எழுத்துணர்தல்" ("real-time character recognition") மற்றும் "நுண்ணறிவு எழுத்துப் உணர்தல்" (intelligent character recognition") என்றும் அழைக்கப்படுகிறது.

பயன்பாடுகள்

எழுத்துணரி பல பயன்பாடுகளுக்குப் பயன்படுத்தப்பட்டுள்ளது. அவற்றில் சில கீழே விளக்கப்பட்டுள்ளன.

1) விலைப்பட்டியல்

விலைப்பட்டியல் படமாக்கம் (imaging) பல வணிகப் பயன்பாடுகளில் நிதி பதிவுகளை கண்காணிக்கவும், பணம் செலுத்துவதைத் தடுக்கவும் பரவலாகப் பயன்படுத்தப்படுகிறது. அரசாங்க நிறுவனங்கள் மற்றும் சுதந்திர நிறுவனங்களில், எழுத்துணரித் தரவு சேகரிப்பு மற்றும் பகுப்பாய்வை எளிதாக்குகிறது. தொழில்நுட்பம் தொடர்ந்து வளர்ச்சியடைந்து வருவதால், எழுத்துணர் தொழில்நுட்பத்திற்காக அதிகமான பயன்பாடுகள் காணப்படுகின்றன, இதில் கையெழுத்து உணர்தலின் பயன்பாடு அதிகம். மேலும், பார்கோடு அங்கீகாரம் போன்ற எழுத்துணரி தொடர்பான பிற தொழில்நுட்பங்கள் சில்லறை மற்றும் பிற தொழில்களில் தினமும் பயன்படுத்தப்படுகின்றன.

2) சட்டத் தொழில்

எழுத்துணரித் தொழில்நுட்பத்தின் பயனாளிகளில் சட்டத் துறையும் ஒன்றாகும். ஆவணங்களை டிஜிட்டல் மயமாக்க எழுத்துணரி பயன்படுத்தப்படுகிறது, மேலும் கணினி

தரவுத்தளத்தில் நேரடியாக நுழைகிறது. சட்ட வல்லுநர்கள் ஒரு சில முக்கிய சொற்களைத் தட்டச்சு செய்வதன் மூலம் பெரிய தரவுத்தளங்களிலிருந்து தேவையான ஆவணங்களைத் தேடலாம்.

3) வங்கி

எழுத்துணரியின் மற்றொரு முக்கியமான பயன்பாடை வங்கியில் காணலாம்; இது மனிதர்களின் ஈடுபாடு இல்லாமல் காசோலைகளைச் செயலாக்க பயன்படுகிறது. ஒரு காசோலையை ஒரு இயந்திரத்தில் செருகலாம், அங்கு கணினி வழங்க வேண்டிய தொகையை ஒளி வருடல் (ஸ்கேன்) செய்து சரியான அளவு பரிமாற்றம் செய்யப்படுகிறது. இந்த தொழில்நுட்பம் கிட்டத்தட்ட அச்சிடப்பட்ட காசோலைகளுக்கு முழுமையாக்கப்பட்டுள்ளது, மேலும் கையால் எழுதப்பட்ட காசோலைகளுக்கு இது மிகவும் துல்லியமானது மற்றும் வங்கிகளில் காத்திருக்கும் நேரத்தைக் குறைக்கிறது.

4) உடல் நலப் பராமரிப்பு

உடல்நலப் பராமரிப்பு காகித வேலைகளைச் செயலாக்க எழுத்துணரித் தொழில்நுட்பத்தைப் பயன்படுத்துவதையும் அதிகரித்துள்ளது. உடல்நலப் பராமரிப்பு வல்லுநர்கள் ஒவ்வொரு நோயாளிக்கும் காப்பீட்டு படிவங்கள் மற்றும் பொது சுகாதார படிவங்கள் உட்பட பெரிய அளவிலான படிவங்களை எப்போதும் கையாள வேண்டும். இந்த எல்லா தகவல்களையும் வைத்திருக்க, தேவையான தரவுகளை மின்னணு தரவுத்தளத்தில் உள்ளிடுவது பயனுள்ளதாக இருக்கும். எழுத்துணரியால் இயக்கப்படும் படிவச் செயலாக்க கருவிகள், படிவங்களிலிருந்து தகவல்களைப் பிரித்தெடுத்து தரவுத்தளங்களில் வைக்க முடியும், இதனால் ஒவ்வொரு நோயாளியின் தரவும் உடனடியாகப் பதிவு செய்யப்படும்.

நிறுவனக் களஞ்சியங்கள் மற்றும் டிஜிட்டல் நூலகங்கள்

5) நிறுவனக் களஞ்சியங்கள்

நிறுவனக் களஞ்சியங்கள் என்பது ஒரு பல்கலைக்கழகம் அல்லது ஆராய்ச்சி நிறுவனத்திற்குள் உருவாக்கப்பட்ட வெளியீடுகளின் மின்னிலக்க/டிஜிட்டல் சேகரிப்புகள் ஆகும். இது ஒரு நிறுவனத்தின் அறிவுசார் தரவுகளின் ஆன்லைன் இடமாகும்; குறிப்பாக ஒரு ஆராய்ச்சி நிறுவனம் சேகரிக்கப்பட்டு, பாதுகாக்கப்பட்டு ஒளிபரப்பப்படுகிறது. இது ஒரு நிறுவனத்தின் வெளியீடுகளைத் திறக்க உதவுகிறது மற்றும் உலகளாவிய அளவில் பார்வை மற்றும் அதிகத் தாக்கத்தை அளிக்கிறது. ஆராய்ச்சிக்கான இடைநிலை அணுகுமுறைகளை இயக்குகிறது மற்றும் ஊக்குவிக்கிறது மற்றும் மின்னிலக்கக் கற்பித்தல் பொருட்கள் (digital teaching materials) மற்றும் உதவிக்கருவிகளின் (aids) வளர்ச்சி மற்றும் பகிர்வுக்கு உதவுகிறது. இது அடிப்படையில் மதிப்பாய்வு செய்யப்பட்ட பத்திரிகைக் கட்டுரைகள், மாநாட்டு நடவடிக்கைகள், ஆராய்ச்சித் தரவு, மோனோகிராஃப்கள், புத்தகங்கள், ஆய்வறிக்கைகள் மற்றும் ஆய்வுக் கட்டுரைகள் மற்றும் விளக்கக்காட்சிகளின் தொகுப்பாகும். அவைகளின் முதல் பங்கு திறந்த அணுகல் இலக்கியங்களை வழங்குவதாகும். இதை நடைமுறையில் செயல்படுத்துவதில் ஆவணங்களை வருடல்/ஸ்கேன் செய்யும் வருடியை/ஸ்கேனரைக் கொண்ட ஒரு அமைப்பை அமைப்பது அடங்கும். இந்த ஸ்கேன்/வருடல் செய்யப்பட்ட ஆவணம் ஆப்டிகல் கேரக்டர் ரெக்னிகிஷன் சிஸ்டத்தின் உள்ளீடாக வழங்கப்படுகிறது, அங்கு தகவல் மின்னிலக்க (டிஜிட்டல்) மயமாக்கப்பட்டுத் தக்கவைக்கப்படுகிறது.

7)ஒழிவழி இசை உணர்தல் (Optical Music Recognition))

தானியங்கி கற்றல் அமைப்பு படங்களிலிருந்து தகவல்களைப் பிரித்தெடுக்கிறது மற்றும் இது முக்கிய ஆராய்ச்சிகளின் ஒரு பகுதியாகும். ஆயிரத்து ஐம்பதுகளில் பிறந்த ஒளிவழி இசை உணர்தல் (Optical music recognition (OMR/ஓஎம்ஆர்)) ஒரு வளர்ந்த

துறையாகும், ஆரம்பத்தில் மின்னணு மற்றும் மின் வேதியியல் முறைகளின் உதவியுடன் இயக்கக்கூடிய வடிவத்தில் திருத்தக்கூடிய அச்சிடப்பட்ட தாள்களை அங்கீகரிப்பதை நோக்கமாகக் கொண்டது. ஒரு ஒளிவழி இசை உணர்தல் (OMR) அமைப்பில் பல்வேறு வகை இசைகளைச் செயலாக்குதல், இயந்திர கற்றல் மற்றும் கணினி சர்வதேச இதழ், தொகுதி. 2, எண் 3, ஜூன் 2012 315 இசை தரவின் பெரிய அளவிலான மின்னிலக்க மயமாக்கல் (digitalization) மற்றும் இசை குறியீட்டில் பன்முகத்தன்மைக்கு இதைப் பயன்படுத்தலாம். படத்தை மேம்படுத்துதல் மற்றும் பிரித்தல் என்பது அடிப்படைப் படியாகும், எனவே காகிதம் அதில் கவனம் செலுத்துகிறது.

8) தானியங்கி எண் தட்டு உணர்தல்

தானியங்கி எண் தகடு உணர்தல் (Automatic Number plate Recognition) வாகனப் பதிவுத் தகடுகளை அடையாளம் காண படங்களில் ஒளிவழி எழுத்து உணர்தலைப் (ஆப்டிகல் கேரக்டர் ரெகனிஷன்) பயன்படுத்துவதற்கான வெகுஜன கண்காணிப்பு நுட்பமாகப் பயன்படுத்தப்படுகிறது. கேமராக்களால் கைப்பற்றப்பட்ட படங்களை உரிமத் தகட்டில் இருந்து கைப்பற்றப்பட்ட எண்கள் உட்பட சேமிக்கவும் ANPR செய்யப்பட்டுள்ளது. ஏ.என்.பி.ஆர் தொழில்நுட்பம் ஒரு பிராந்திய குறிப்பிட்ட தொழில்நுட்பமாக இருப்பதால் இடத்திலிருந்து இடத்திற்கு மாறுபடும். அவை பல்வேறு காவலர் படையினரால் பயன்படுத்தப்படுகின்றன மற்றும் பயன்பாட்டுக்குச் செல்லும் சாலைகளில் மின்னணு கட்டண வசூல் மற்றும் போக்குவரத்து அல்லது தனிநபர்களின் நகர்வுகளை பட்டியலிடுகின்றன.

9) கையெழுத்து உணர்தல்

கையெழுத்து உணர்தல் என்பது காகித ஆவணங்கள், புகைப்படங்கள், தொடுதிரைகள் மற்றும் பிற சாதனங்கள் போன்ற மூலங்களிலிருந்து உணர்ந்து கொள்ளக்கூடிய கையால் எழுதப்பட்ட உள்ளீட்டைப் பெற்று விளக்கும் ஒரு கணினியின்

திறன் ஆகும். எழுதப்பட்ட உரையின் படம் ஒளிவழி வருடல் (ஆப்டிகல் ஸ்கேனிங்) (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) அல்லது நுண்ணறிவுச் சொல் உணர்தல் (intelligent word recognition) மூலம் ஒரு காகிதத்திலிருந்து "ஆஃப் லைன்" உணரப்படலாம். மாற்றாக, பேனா முனையின் இயக்கங்கள் "வரியில்" உணரப்படலாம், எடுத்துக்காட்டாக பேனா அடிப்படையிலான கணினித் திரை மேற்பரப்பு.

10) முதியோர் மற்றும் பார்வையற்றோரின் வாழ்க்கையை ஆதரித்தல்

ஆயிரத்து எழுபதுகளில் அமெரிக்காவின் குர்ஸ்வீல் கணிபொறி தயாரிப்புகள் இன்க் (Kurzweil Computer Products Inc) உலகின் முதல் ஆம்னி எழுத்துரு முறையை (Omni font system) உருவாக்கியது. எழுத்துணர்தல் (OCR) மென்பொருளால் இந்த எழுத்துருவை அடையாளம் காண முடியும். உடனடியாக, எழுத்துணர் (ஓ.சி.ஆர்) தொழில்நுட்பம் பேச்சுத் தொகுப்புத் தொழில்நுட்பத்துடன் ஒருங்கிணைக்கப்பட்டு, உரையை வாசிக்கும் மற்றும் புரிந்துகொள்ளும் திறன் கொண்டது.

வேறு வார்த்தைகளில் கூறுவதானால், உரை எழுத்துணர்தல் (OCR) மென்பொருளால் குறியத்திறவு (டிகோட்) செய்யப்படுவது மட்டுமல்லாமல், பேச்சு தொகுப்பு இயந்திரத்தால் படிக்கப்படுகிறது. கணினிமயமாக்கப்பட்ட குரல் உரைகள், செய்தித்தாள்கள், முதியவர்கள் மற்றும் பார்வையற்றோருக்கான பத்திரிகைகளில் உரை வாசிப்பதற்குப் பயன்படுத்தப்பட்டு, அவர்களின் வாழ்க்கையை எளிதாக்குகிறது.

11) சட்ட நிறுவனங்கள் மற்றும் நீதிமன்றங்களில் ஆவணங்களை ஏற்பாடு செய்தல்

ஒவ்வொரு வழக்கிலும், சட்ட ஆவணங்கள், பதிவுகள் ஏராளமானவை மற்றும் சிக்கலானவை. எந்தவொரு முக்கியமான விவரங்களையும் ஆவணங்களையும்

தவறவிடாமல் பார்த்துக் கொள்ள, வழக்கறிஞர்கள் ஒழுங்கமைத்து நீண்ட நேரம் தேட வேண்டும்.

ஒளிவழி எழுத்துணர்தல் (OCR) மென்பொருளுக்கு நன்றி, வழக்கறிஞர்கள் அனைத்து ஆவணங்களையும் மிக விரைவாக மின்னிலக்க மயமாக்க (digitalization) முடியும். தேவைப்படும்போது, முக்கிய சொல், தேதி, கோப்பு, பெயர், எளிய, வசதியான மற்றும் விஞ்ஞான வழியில் ஆவணங்களை எளிதாகக் கண்டுபிடிப்பார்கள். இதன் பொருள், வழக்கறிஞர்கள் கையேடு பணிகளைச் செய்ய பல உதவியாளர்களை நியமிக்க வேண்டியதில்லை, ஆனால் பணியைக் கண்காணிப்பதை உறுதிசெய்கிறார்கள்.

12) மதிப்புமிக்க ஆவணங்களைப் பாதுகாத்தல்

பண்டைய நூலகங்கள், வரலாற்று-கலாச்சார மையங்கள் அல்லது அருங்காட்சியகங்கள்... நிறைய கையெழுத்துப் பிரதிகள், ஆவணங்கள், நினைவுக் குறிப்புகளைச் சேமித்து வைக்கும் இடங்கள்... இந்த ஆவணங்களைச் சேமித்து பாதுகாக்கும் செயல்முறை எளிதானது அல்ல. அவை காலப்போக்கில் கரையான்கள் மற்றும் சேதங்களுக்கு ஆளாகின்றன. ஒரு பெரிய அளவிலான உரையின் துல்லியமான மற்றும் முழு கையேடு உள்ளீடு மிகவும் கடினம் மற்றும் கடினமானது, இது பல தசாப்தங்களாக ஆகலாம்.

இருப்பினும், பல நிறுவனங்களுக்கு அந்த சிக்கலை எளிமையான முறையில் தீர்க்க உதவும் வகையில் OCR தொழில்நுட்பம் பிறந்தது. முக்கியமான ஆவணங்கள் காகித வடிவங்களிலிருந்து மென்மையான கோப்புகளாக மாற்றப்பட்டுள்ளன, இதனால் பல இலக்கிய மரபுகளைச் சேமித்து வைப்பது எளிதாகிறது.

13) தனிப்பட்ட அடையாளம்

வங்கிக் கணக்கிற்குப் பதிவு செய்யும்போது, உறுப்பினர் அட்டை அல்லது அடையாள சரிபார்ப்பு தேவைப்படும் வேறு ஏதேனும் செயல்களைத் திறக்கும்போது, நீங்கள் கூட்டாளர்களுக்கு துல்லியமான மற்றும் முழுமையான தனிப்பட்ட தகவல்களை வழங்க வேண்டும். அறிவிப்பு செயல்முறை உங்களுக்கும் சேவை வழங்குநருக்கும் சிறிது நேரம் எடுக்கும். சில நேரங்களில், தவறான தரவு உள்ளீடு பின்னர் தேவையற்ற சிக்கல்களுக்கு வழிவகுக்கிறது.

OCR மென்பொருளுடன், அடையாள அட்டைகள், பாஸ்போர்ட், ஓட்டுநர் உரிமங்கள் மற்றும் பல ஆவணங்கள் போன்ற சட்ட ஆவணங்களை நோட்டரி அலுவலகங்கள், காவலர் அலுவலகங்கள், விமான நிலையங்கள் மற்றும் தனிப்பட்ட தகவல்களை செயலாக்க வேண்டிய பல நிறுவனங்கள் மற்றும் சேவைகள் மூலம் விரைவாக வருடல்/ஸ்கேன் செய்யலாம். அதிகமான எழுத்துக்கள் இல்லாத தனிப்பட்ட ஆவணங்களுக்கு, எழுத்துணரி (OCR) தொழில்நுட்பம் கிட்டத்தட்ட முழுமையான, உள்ளீட்டு பிழைகளைக் குறைக்கும் திறனைத் துல்லியமாக அடையாளம் காணும் திறனைக் கொண்டுள்ளது. மேலும், இயந்திரங்கள் மூலம் தகவல்களை அடையாளம் கண்டு பிரித்தெடுப்பது ஏஜென்சிகள், பிரிவுகள் மற்றும் நிறுவனங்கள் தகவல்களை எளிதாக சேமிக்க உதவுகிறது மற்றும் எந்த நேரத்திலும் பயனர் தகவல்களை மீட்டெடுக்க முடியும்.

14) விலைப்பட்டியல் மற்றும் பல வகையான ஆவணங்களை செயலாக்குதல்

அனைத்து ஏஜென்சிகள் மற்றும் நிறுவனங்கள் ஆயிரக்கணக்கான ஆவணங்கள் மற்றும் ஆவணங்களைக் கொண்டுள்ளன, காகிதம், PDF, JPG போன்றவற்றில் அச்சிடப்பட்ட கையால் எழுதப்பட்ட ஆவணங்கள் போன்ற பல வடிவங்கள் உள்ளன. அந்த தரவு கணினியில் கிடைக்கிறது, அல்லது அந்த முடிவற்ற அளவை

செயலாக்க அதிக நேரம் எடுக்கும்; மேலும், தரவு உள்ளீட்டில் பிழைகள் நிகழ்தகவு மிகப் பெரியது.

பல ஏஜென்சிகள் மற்றும் நிறுவனங்கள் ஒப்பந்தங்கள், விலைப்பட்டியல், செலவுச் சீட்டுகள் மற்றும் பல ஆவணங்களை மின்னிலக்க (டிஜிட்டல்) ஆவணங்களாக மாற்ற தேர்வு செய்கின்றன, அந்தத் தரவை நிதி அறிக்கைகளுக்கு எளிதாகப் பயன்படுத்துவதற்கும், ஆவணங்களைச் சேமிப்பதற்கும் அல்லது பரிமாறிக்கொள்வதற்கும் எழுத்துணரி மென்பொருள் மேலும் ஒரு சிறந்த தேர்வாகும்.

தற்போது, உலகெங்கிலும் உள்ள அறுபது விழுக்காட்டுக்கும் மேற்பட்ட பெரிய நிறுவனங்கள் வணிக செயல்முறைகளில் பல படிகளுக்குத் தரவை உள்ளிட எழுத்துணரியைப் (OCR) பயன்படுத்துகின்றன. எழுத்துணர் தொழில்நுட்ப நிறுவனங்களைத் தானாகக் கணினியில் தரவைச் சேமிக்க, மின்னஞ்சல், தொலைநகல் அல்லது பாரம்பரிய EDI போன்ற மற்றொரு தளத்துடன் எளிதாக ஒருங்கிணைக்க, மாற்ற அல்லது இணைக்க அனுமதிக்கிறது.

நிறுவனங்கள் மற்றும் பயனர்கள் இருவருக்கும் ஒரே மாதிரியான பயன்பாடுகளால் எழுத்துணரி (OCR) தொழில்நுட்பம் மேலும்மேலும் பிரபலமாகி வருகிறது. எதிர்காலத்தில், பல மனித நடவடிக்கைகளுக்குத் திருப்புமுனைத் தயாரிப்புகளை உருவாக்க, எழுத்துணரி (OCR) பல மேம்பட்ட தொழில்நுட்பங்களுடன் ஒருங்கிணைக்கப்படும்.

வியட்நாமில், சந்தையில் மிகத் துல்லியத்துடன் ஒருங்கிணைந்த எழுத்துணரி (OCR) தொழில்நுட்பத்தைக் கொண்ட தயாரிப்புகளில் ஒன்று FPT கார்ப்பரேஷனின் கீழ் FPT தொழில்நுட்ப கண்டுபிடிப்புத் துறையால் உருவாக்கப்பட்ட FPT.AI Vision ஆகும். அடையாள அட்டை, ஓட்டுநர் உரிமம், பாஸ்போர்ட் போன்ற தனிப்பட்ட ஆவணங்களின்

தகவல்களை 98% வரை துல்லியத்துடன் கண்டறிந்து பிரித்தெடுப்பதற்கான தீர்வாகும், இது சந்தையை அறிதல்/புரிடல் தரத்தில் வழிநடத்துகிறது.

கூடுதலாக, FPT.AIஇன் ஒளிவழி எழுத்துணரி (OCR) தொழில்நுட்பம் ஒவ்வொரு கூட்டாளியின் தேவைகளைப் பொறுத்து விலைப்பட்டியல், ஒப்பந்தங்கள் மற்றும் பல குறிப்பிட்ட தேவைகள் போன்ற ஆவணங்களின் தகவல்களை அடையாளம் கண்டு பிரித்தெடுக்கும் திறனைக் கொண்டுள்ளது. FPT.AI விஷன் வணிகங்களை ஆவணங்களை டிஜிட்டல் மயமாக்கவும் வாடிக்கையாளர் தகவல்களை விரைவாக அடையாளம் காணவும் உதவுகிறது. எழுத்துணரிக்கு நன்றி, தரவு நுழைவு நேரம் சுருக்கப்பட்டது, தகவல் மிகவும் துல்லியமானது, நேரம், முயற்சி மற்றும் இயக்கச் செலவுகளை மிச்சப்படுத்துகிறது.

15) ஒளிவழி எழுத்துணர்தல் (OCR) - புத்திசாலித்தனமான எதிர்காலத்துடன் முதிர்ந்த தொழில்நுட்பம்

மின்னணு ஆவண மேலாண்மை அமைப்புகளில் மறுநிகழ்வின் பங்கைப் பாருங்கள். ஓ.சி.ஆர் தொழில்நுட்பம் ஒப்பீட்டளவில் புதியது என்று பலர் நினைக்கிறார்கள், இது சில ஆண்டுகளாக நம்மிடம் உள்ளது. இருப்பினும், ஓ.சி.ஆரின் வேர்களை நீண்ட தூரம் காணலாம். உண்மையில், எழுத்துணர்தல் (OCR) சம்பந்தப்பட்ட முதல் காப்புரிமைகள் ஆயிரத்து எண்ணுற்று ஒன்பதில் வழங்கப்பட்டன. ஆயிரத்து எண்ணுற்று எழுபதாம் ஆண்டில், போஸ்டனைச் சேர்ந்த ஒரு திரு. கேரி (Mr. Carey from Boston) ஒரு பட பரிமாற்ற முறைக்குக் காப்புரிமை பெற்றார், இது ஒளிச்சேர்க்கைகளின் மொசைக் பயன்படுத்தியது. இது ஒரு "விழித்திரை ஸ்கேனரின்" ஆரம்ப எடுத்துக்காட்டு. தேசிய பணியகத்தின் கண்டுபிடிப்பாளரும் விஞ்ஞானியுமான ஜேக்கப் ராபினோவின் கூற்றுப்படி, "வாஷிங்டனில் நுண்ணறிவு இயந்திர ஆராய்ச்சி கழகத்தை நிறுவிய OCR ஆர்ட்வாஸில் சிறந்த அமெரிக்க முன்னோடி டேவிட் ஷெப்பர்ட் (David Shepard) ஒளிவழி எழுத்து

அறிதல் கருவியை (OCR equipment) உருவாக்க மற்றும் உருவாக்க ஆயிரத்து தொள்ளாயிரத்து ஐம்பதில் டி.சி. திரு. ஷெப்பர்ட் தற்போது அறிவாற்றல் இமேஜிங் சிஸ்டம்ஸ் வாரியத்தின் தலைவராக உள்ளார், இன்னும் வணிகத்தில் தீவிரமாக உள்ளார். ஆயிரத்து தொள்ளாயிரத்து ஐம்பதி ஒன்றாம் ஆண்டில் முதல் நவீன நடைமுறை எழுத்துணர் (ஓ.சி.ஆர்) முறையை உருவாக்கி காப்புரிமை பெற்ற பெருமைக்குரியவர் ஹெய்ஸ். அதைத் தொடர்ந்து வந்த தசாப்தங்களில், புதிய நிறுவனங்கள் தானியங்கித் தரவு பிடிப்புத் துறையை ஆதரித்தன. ஒளிவழி எழுத்துணர்தல் (OCR) கருவிகளின் பெரிய அளவிலான வணிக வளர்ச்சி ஆயிரத்து தொள்ளாயிரத்து எழுபதுகளில் தொடங்கியது. கடந்த ஆண்டுகளில், ஆப்டிகல் கேரக்டர் அங்கீகாரம் அமைப்புகள் ஒரு வகையான சிறப்பு நோக்க வாசகர்களிடமிருந்து இன்றைய பல்நோக்கு உற்பத்தி மற்றும் ஊடாடும் ஆன்லைன் அமைப்புகள் வரை நீண்ட தூரம் வந்துள்ளன. இந்த முன்னேற்றம் தரவுப் பிடிப்புச் செலவுகளை (data capture costs) குறைத்து, மேலும் நம்பகமான மற்றும் துல்லியமான ஒளிவழி எழுத்துணர்தல் (OCR) ஒழுங்குமுறைகளின் வளர்ச்சியை ஏற்படுத்தியுள்ளது.

முடிவுரை

எழுத்துணரி (OCR) ஒளிவழி எழுத்துணர்தல் (Optical Character Recognition) என்றும் அழைக்கப்படுகிறது. எழுத்தை வாசிப்பான் என்பது ஒரு முழுமையான ஒழுங்குமுறை ஆகும். அச்சிடப்பட்ட அல்லது கையால் எழுதப்பட்ட எண்ணெழுத்து படிவத்தை வருடுதல் (ஸ்கேன்) மூலம் மின்னணு வேகத்தில் எழுத்துக்களைப் படித்து அறியும். மிகச் சமீபத்தில், நுண்ணறிவு எழுத்து உணர்தல் (Intelligent Character Recognition (ICR) ஐ.சி.ஆர்) என்ற சொல் விளக்கும் செயல்முறையை விவரிக்க பயன்படுத்தப்பட்டது. படத் தரவு, குறிப்பாக எண்ணெழுத்து உரை, எழுத்துக்கள்

படங்களைக் கொண்ட எழுத்துணரி படிவங்களின் செயல்பாடு ஸ்கேன்/வருடல் செய்யலாம் ஸ்கேனர் மற்றும் எழுத்துணரியின் உணர்தல் இயந்திரம் மூலம் கணினிப் படங்களை விளக்குகிறது மற்றும் கையால் எழுதப்பட்ட படங்களை திருப்புகிறது அல்லது அச்சிடப்பட்ட எழுத்துக்கள் ASCII தரவுகளில் (இயந்திரம் படிக்கக்கூடியவை எழுத்துக்கள்). எனவே, தரவை விரைவாகத் தானியக்கமாக்க எழுத்துணரி பயனர்களை அனுமதிக்கிறது படிவங்களிலிருந்து பிடிக்கவும், தரவைக் குறைக்க விசை அழுத்தங்களை அகற்றவும் நுழைவுச் செலவுகள் மற்றும் அதிக அளவு துல்லியத்தைப் பராமரிக்கின்ற படிவங்கள் செயலாக்கப் பயன்பாடுகளில் தேவை. எழுத்துணரியின் அம்சங்கள் தொழில்நுட்பம் ஒரு முழுமையான படிவ செயலாக்கத்தை வழங்குகிறது ஆவணங்கள் தீர்வு. வழக்கமாக, எழுத்துணரி ஒரு பயன்படுத்துகிறது மட்டு கட்டமைப்பு திறந்த, அளவிடக்கூடிய மற்றும் பணிப்பாய்வு கட்டுப்படுத்தப்படுகிறது. இதில் படிவங்கள் வரையறை, ஸ்கேனிங், படம் ஆகியவை அடங்கும் முன் செயலாக்கம் மற்றும் உணர் திறன்கள். அறிவுசார் எழுத்து அறிவான் (Intelligent Character Recognition (ICR/ஐ.சி.ஆர்) நுண்ணறிவு எழுத்து அங்கீகாரம் (ஐ.சி.ஆர்) என்றால் என்ன? கையால் எழுதப்பட்ட படங்களை மாற்றும் திறன் கொண்ட எழுத்துணரி அல்லது ASCII தரவுகளில் அச்சிடப்பட்ட எழுத்துக்கள் சில நேரங்களில் எழுத்துணரி ஆகும் ஐ.சி.ஆர் என அழைக்கப்படுகிறது. ஐ.சி.ஆரின் வித்தியாசம், எழுத்துணரி மற்றும் குறி உணரி (Optical Mark Reader (OMR) ஐ.சி.ஆர் மற்றும் ஓ.சி.ஆர் ஆகியவை இமேஜிங்கில் பயன்படுத்தப்படும் அங்கீகார இயந்திரங்கள்; OMR என்பது தரவு சேகரிப்பு தொழில்நுட்பமாகும். அங்கீகார இயந்திரம் தேவை. எனவே, அடிப்படையில் ஓ.எம்.ஆர் கையால் அச்சிடப்பட்ட அல்லது இயந்திரத்தால் அச்சிடப்பட்டதை அடையாளம் காண முடியாது எழுத்துக்கள். இருப்பினும், OCR

தொழில்நுட்பத்தில், பதில் “டிக்” அல்லது “மார்க்” இல் உள்ள கேள்வி எழுத்துணரி என்றும் அழைக்கப்படுகிறது.

துணை நூல்கள்

Baird H.S. & R. Fossey. 1991. A 100-Font Classifier. Proceedings ICDAR-91, Vol. 1, p. 332-340, 1991.

"Basic OCR in OpenCV | Damiles". Blog.damiles.com. November 20, 2008. Retrieved on 11.10.2020.

"Breaking a Visual CAPTCHA". Cs.sfu.ca. December 10, 2002. Retrieved 11.10.2020.

Caillot, J-P. Review of OCR Techniques. NR-note, BILD/08/087.

"Code and Data to evaluate OCR accuracy, originally from UNLV/ISRI". Google Code Archive.

"Extracting text from images using OCR on Android". June 27, 2015. Archived from the original on 13.10.2020.

Fehr, Tiff, How We Sped Through 900 Pages of Cohen Documents in Under 10 Minutes, Times Insider, The New York Times, March 26, 2019.

Govindan V. K. & A.P. Shivaprasad. Character Recognition - a Review. Pattern Recognition, Vol. 23, No &, P. 671-683, 1990.

Gupta, Maya R.; Jacobson, Nathaniel P.; Garcia, Eric K. 2007. "OCR binarisation and image pre-processing for searching historical documents" (PDF). Pattern Recognition. 40 (2): 389. doi:10.1016/j.patcog.2006.04.043. Archived from the original (PDF) on October 16, 2015. 10.10.2020.

Holley, Rose (April 2009). "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". D-Lib Magazine. Retrieved 12.10.2020.

"How does OCR document scanning work?". Explain that Stuff. January 30, 2012. Retrieved 14.10.2020.

"How to optimize results from the OCR API when extracting text from an image? - Haven OnDemand Developer Community". Archived from the original on 12.10.2020.

"How To Crack Captchas". andrewt.net. June 28, 2006. Retrieved 10.10.2020.

"How OCR Software Works". OCRWizard. Retrieved 12.10.2020.

Impedovo, S. & L. Ottaviano & S. Occhinegro. 1991. Optical Character Recognition - A survey. Int. Journal of PRAI, Vol. 5, No 1& 2, p. 1-24, 1991.

Milyaev, Sergey; Barinova, Olga; Novikova, Tatiana; Kohli, Pushmeet; Lempitsky, Victor 2013. "Image binarisation for end-to-end text understanding in natural images" (PDF). Document Analysis and Recognition (ICDAR) 2013. 12th International Conference on: 128–132. doi:10.1109/ICDAR.2013.33. ISBN 978-0-7695-4999-6. S2CID 8947361. Retrieved 13.10.2020.

Mori S.; C.Y. Suen & K. Yamamoto. 1992. Historical Review of OCR research and Development. IEEE Proceedings, special issue on OCR, p. 1029-1057, July 1992.

Nagy. G. 1992. At the Frontiers of OCR. IEEE Proceedings, special issue on OCR, p.1093-1100, July 1992.

Optical Character Rcognition. From Wikipedia, the free encyclopedia. Downloaded on 10.10.2020.

"OCR Introduction". Dataid.com. Retrieved 14.10.2020.

"Optical Character Recognition (OCR) – How it works". Nicomsoft.com. Retrieved 11.10.2020.

Plamondon R. & G. Lorette. 1989. Automatic Signature Verification and Writer Identification -The State of the Art. Pattern Recognition, Vol. 22,No 2, p. 107-131, 1989.

Pati, P.B.; Ramakrishnan, A.G. (May 29, 1987). "Word Level Multi-script Identification". Pattern Recognition Letters. 29 (9): 1218–1229. doi:10.1016/j.patrec.2008.01.027.

Pavlidis, T. 1993. Recognition of printed text under realistic conditions. Pattern Recognition Letters 14, p. 317-326, 1993.

Resig, John (January 23, 2009). "John Resig – OCR and Neural Nets in JavaScript". Ejohn.org. Retrieved 13.10.2020.

Riedl, C.; Zanibbi, R.; Hearst, M. A.; Zhu, S.; Menietti, M.; Crusan, J.; Metelsky, I.; Lakhani, K. (February 20, 2016). "Detecting Figures and Part Labels in Patents: Competition-Based Development of Image Processing Algorithms". International Journal on Document Analysis and Recognition. 19 (2): 155. arXiv:1410.6751. doi:10.1007/s10032-016-0260-8. S2CID 11873638.

Sezgin, Mehmet; Sankur, Bulent. 2004. "Survey over image thresholding techniques and quantitative performance evaluation" (PDF). Journal of Electronic Imaging. 13 (1): 146.

Bibcode:2004JEI....13..146S. doi:10.1117/1.1631315. Archived from the original (PDF) on October 16, 2015. Retrieved 10.10.2020.

Schantz, H. F. The History of OCR. Recognition Technology Users Association, VT, 1982. "The History of OCR". Data Processing Magazine. 12: 46. 1970.

Scurmann, J. Reading Machines. Proceedings IJCPR, Munich, p. 1031-1044, 1982.

Suen, C.Y.; Plamondon, R.; Tappert, A.; Thomassen, A.; Ward, J.R.; Yamamoto, K. (May 29, 1987). Future Challenges in Handwriting and Computer Applications. 3rd International Symposium on Handwriting and Computer Applications, Montreal, May 29, 1987. Retrieved on 12.10.2020.

Smith, Ray. 2007. "An Overview of the Tesseract OCR Engine" (PDF). Retrieved on 13.10.2020.

Suen, C.Y. ; M. Berthod & S. Mori. 1980. Automatic Recognition of Handprinted Characters - The State of the Art. IEEE Proceedings, Vol. 68, No. 4, p.469-487, April 1980

Sarantos Kapidakis, Cezary Mazurek, Marcin Werla. 2015. Research and Advanced Technology for Digital Libraries. Springer. p. 257. ISBN 9783319245928. Retrieved on 12.10.2020.

Tappert, C. C.; Suen, C. Y.; Wakahara, T. 1990. "The state of the art in online handwriting recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 12 (8): 787. doi:10.1109/34.57669. S2CID 42920826.

"The basic pattern recognition and classification with openCV | Damiles". Blog.damiles.com. November 14, 2008. Retrieved on 14.10.2020.

Trier, Oeivind Due; Jain, Anil K. 1995. "Goal-directed evaluation of binarisation methods" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 17 (12): 1191–1201. doi:10.1109/34.476511. Retrieved 13.10.2020.

"What is the point of an online interactive OCR text editor? - Fenno-Ugrica". February 21, 2014.

Young T. Y. & K-S Fu. 1986. Handbook of Pattern Recognition and Image Processing. Academic Press, 1986.

வினாவங்கி

1. பொருத்தமான விடையைத் தேர்ந்தெடுத்தல்

1) எழுத்து உணர்தலின் வேறுபட்ட களங்களை அடிப்படையில் எவ்வாறு பிரிக்கலாம்?

அ. தனித்தனிஎழுத்துகள், சேர்ந்து எழுதப்பட்ட எழுத்துகள் என இரண்டாகப் பிரிக்கலாம்.

ஆ. அச்சடிக்கப்பட்டது, எழுத்தப்பட்டது இரண்டாகப் பிரிக்கலாம்.

இ. உணர்தல், சரிபார்த்தல் என இரண்டாகப் பிரிக்கலாம்.

ஈ. ஆஃப்லைன், ஆண்லைன் என இரண்டாகப் பிர்க்கலாம்.√.

2) ஆஃப் லைன் எழுத்துணர்தலை எவ்வாறு பிரிக்கலாம்?

அ. தனித்தனிஎழுத்துகள், சேர்ந்து எழுதப்பட்ட எழுத்துகள் என இரண்டாகப் பிரிக்கலாம்.

ஆ. அச்சடிக்கப்பட்டது, எழுத்தப்பட்டது இரண்டாகப் பிரிக்கலாம்.

இ. உணர்தல், சரிபார்த்தல் என இரண்டாகப் பிரிக்கலாம்.

ஈ. உச்சரிக்கப்பட்டது, உச்சரிக்கப்படாதது என இரண்டாகப் பிரிக்கலாம்.

3) எழுத்து உணர்தலின் தோற்றம் உண்மையில் ------- ஆண்டில் காணப்படுகின்றது.

அ. 1860ஆம்

ஆ. 1870ஆம்√

இ. 1880ஆம்

ஈ. 1890ஆம்

4) விழித்திரை ஸ்கேனரை கண்டுபிடித்தவர்.

அ. சி.ஆர். கேசரி√

ஆ. சி.ஆர். கிரிகரி

இ. சி.ஆர். நரகரி

ஈ. சி.ஆர். புகாரி

5) இரண்டு தசாப்தங்களுக்குப் பிறகு போலந்தைச் சார்ந்த ----- தொடர்ச்சியான ஸ்கேனரைக் கண்டுபிடித்தார்.

அ. பி. கிப்போ

ஆ. பி. நிப்கோ√

இ. பி. பெப்போ

ஈ. பி. டெப்போ

6) ------- எழுத்துணரி உடனான சோதனைகள் மூலம் பார்வையற்றவர்களுக்கு உதவ சாதனங்களை உருவாக்க பல முயற்சிகள் மேற்கொள்ளப்பட்டன..

அ. 16ஆம் நூற்றாண்டின் முதல் தசாப்தங்களில்.

ஆ. 17ஆம் நூற்றாண்டின் முதல் தசாப்தங்களில்.

இ.18ஆம் நூற்றாண்டின் முதல் தசாப்தங்களில்

ஈ. 19ஆம் நூற்றாண்டின் முதல் தசாப்தங்களில்√

7) டிஜிட்டல் கணினியின் வளர்ச்சியுடன் ------- எழுத்துணரியின் நவீன பதிப்பு தோன்றவில்லை.

அ. 1940களின் நடுப்பகுதி வரை√

ஆ. 1950களின் நடுப்பகுதி வரை

இ. 1960களின் நடுப்பகுதி வரை

ஈ. 1970களின் நடுப்பகுதி வரை

8) ------ தொழில்நுட்ப புரட்சி அதிவேகமாக முன்னேறி வந்தது,

அ. 1930வாக்கில்

ஆ. 1940வாக்கில்

இ. 1950வாக்கில்√

ஈ. 1960வாக்கில்

9) ------- நடுப்பகுதியில் எழுத்துணரும் இயந்திரங்கள் வணிக ரீதியாக மாறிக் கிடைத்தது.

அ. 1940களின்

ஆ. 1950களின் √

இ. 1960களின்

ஈ. 1970களின்

10) முதல் உண்மையான எழுத்துணரி வாசிப்பு இயந்திரம் 1954இல் --------

அ. லேனர்ஸ் டைஜெஸ்டில் நிறுவப்பட்டது.

ஆ. இந்துஸ்தான் டைம்ஸில் நிறுவப்பட்டது.

இ. ரீடர்ஸ் டைஜெஸ்டில் நிறுவப்பட்டது.√

ஈ. அமேரிக்கன் டைம்ஸில் நிறுவப்பட்டது.

11) 1960 முதல் 1965 வரையிலான காலகட்டத்தில் தோன்றிய வணிக எழுத்துணரி ஒழுங்குமுறைகளை எழுத்துணரியின் ------ என்று அழைக்கலாம்.

அ. நான்காம் தலைமுறை

ஆ. மூன்றாம் தலைமுறை

இ. இரண்டாம் தலைமுறை

ஈ. முதல் தலைமுறை√

12) ------ 1965ஆம் ஆண்டில் நியூயார்க்கில் நடந்த உலக கண்காட்சியில் காட்சிக்கு வைக்கப்பட்டது.

அ. ஐபிஎம் 1287 √

ஆ. ஐபிஎம் 1288

இ. ஐபிஎம் 1289

ஈ. ஐபிஎம் 1290

13) -------- எழுத்துணரி தேவைகள் பற்றிய முழுமையான ஆய்வு முடிக்கப்பட்டது.

அ. 1965ஆம் ஆண்டில்

ஆ. 1966ஆம் ஆண்டில்√

இ. 1966ஆம் ஆண்டில்

ஈ. 1966ஆம் ஆண்டில்

14) 1970களின் நடுப்பகுதியில் தோன்றிய -------- அமைப்புகளுக்குச் சவாலானது மோசமான தரம் மற்றும் பெரிய அச்சிடப்பட்ட மற்றும் கையால் எழுதப்பட்ட எழுத்துத் தொகுப்புகளின் ஆவணங்கள் ஆகும்.

அ. ஒன்றாம் தலைமுறை எழுத்துணரி√

ஆ. இரண்டாம் தலைமுறை எழுத்துணரி

இ. மூன்றாம் தலைமுறை எழுத்துணரி

ஈ. நான்காம் தலைமுறை எழுத்துணரி

15) 1914ஆம் ஆண்டில் ---------- ஒரு இயந்திரத்தை உருவாக்கி, எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றினார் (Dhavale, 2017).

அ. மானுவல் கோல்ட்பர்க்

ஆ. இமானுவேல் கோல்ட்பர்க் √

இ. மானுவேல் கால்ட்வெல்

ஈ. இமானுவேல் கால்ட்வெல்

2. பொருத்துக (4 வினா-விடை தொகுப்பு அடங்கிய 10 வினாக்கள்)

1) பொருத்துக

அ. அறிவு அடிப்படையிலான எழுத்துணரி - 1954இல்

ஆ. எழுத்து உணர்தலின் தோற்றம் உண்மையில் - 1950வாக்கில்

இ. தொழில்நுட்ப புரட்சி அதிவேகமாக முன்னேறி வந்தது- 1870ஆம் ஆண்டில்

ஈ. முதல் உண்மையான எழுத்துணரி வாசிப்பு இயந்திரம் 1 ரீடர்ஸ் டைஜெஸ்டில் (Reader's Digest) நிறுவப்பட்டது - 2019 ஆண்டில்

2) எழுத்துணையின் முன்னேற்றம்

அ. 1870 - செயற்கை அறிவு அடிப்படையிலான எழுத்துணரி

ஆ. 1940 - முதல் எழுத்துணரி இயந்திரத்தின் தோற்றம்

இ. 1950 - எழுத்துணரியின் தற்கால பதிப்பு

ஈ. 2019 - முதல் முயற்சிகள்

3) கண்டுபிடித்தவரும் கண்டுப்பும்

அ. சி.ஆர். கேரி (C.R.Carey) - கையடக்க ஸ்கேனரான ஆப்டோஃபோன் (Optophone)

ஆ. பி. நிப்கோ - எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றும் இயந்திரத்திரம்

இ. இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg)- தொடர்ச்சியான ஸ்கேனர்

ஈ. எட்மண்ட் ஃபோர்னியர் டி ஆல்பே (Edmund Fournier d'Albe) - விழித்திரை ஸ்கேனர் (retina scanner)

4) எழுத்துணரியின் முன்னேற்றம்- காலகட்டத்தை தலைமுறையுடன் பொருத்துக

அ. 1960 - 1965 - மக்களுக்கான எழுத்துணரி

ஆ. 1965 - 1975 - மூன்றாம் தலைமுறை எழுத்துணரி

இ. 1975 - 1985 - இரண்டாம் தலைமுறை எழுத்துணரி

ஈ. 1986 -> - முதல் தலைமுறை எழுத்துணரி

5) ஆண்டையும் கண்டுபிடிப்பையும் பொருத்துக

அ. இமானுவேல் கோல்ட்பர்க் (Emanuel Goldberg) ஒரு இயந்திரத்தை உருவாக்கி, எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றினார் - 1974ஆம் ஆண்டில்

ஆ. இமானுவேல் கோல்ட்பர்க்"புள்ளிவிவர இயந்திரம்" என்று அழைக்கப்பட்ட ஒன்றை உருவாக்கினார் - 1931ஆம் ஆண்டில்

இ. அவரது கண்டுபிடிப்புக்காக அமெரிக்காவின் காப்புரிமை எண் 1,838,389 வழங்கப்பட்டது. காப்புரிமையை ஐ.பி.எம்.-ஆல் பெறப்பட்டது. - 1920களின் பிற்பகுதியிலும் 1930களில்

ஈ. ரே குர்ஸ்வீல் (Ray Kurzweil) குர்ஸ்வீல் கம்ப்யூட்டர் தயாரிப்புகள், இன்க் (Kurzweil Computer Products, Inc.) என்ற நிறுவனத்தைத் தொடங்கினார் - 1914ஆம் ஆண்டு

6) பொருத்துக

அ. கியூனிஃபார்ம் (Cuneiform) மற்றும் டெசராக்ட் (Tesseract) போன்ற மென்பொருள்கள் - ஒரு படத்தை பிக்சல்-பை-பிக்சல் அடிப்படையில் சேமிக்கப்பட்ட கிளிஃபுடன் ஒப்பிடுவதை உள்ளடக்குகிறது;

ஆ. நவீன ஒளிமூலம் எழுத்துப் புரிதல் (OCR) மென்பொருளானது - தேர்வுக்குரிய எழுத்துக்களின் தரவரிசைப் பட்டியலை உருவாக்கக்கூடும்

இ. மைய ஒளிமூலம் எழுத்துப் புரிவான் (கோர் ஓ.சி.ஆர்.) வழிமுறையின் இரண்டு அடிப்படை வகைகள் உள்ளன, அவை – ஓசிஆர்ஓபஸ் OCRopus அல்லது தெஸ்ஸெராக்ட் (Tesseract) போன்ற நரம்பியல் வலையமைப்புகளைப் (நெட்வொர்க்குகளைப்) பயன்படுத்துகிறது

ஈ. மேட்ரிக்ஸ் பொருத்தம் (Matrix matching) என்பது - எழுத்துப் புரிதலுக்கு (character recognition) இரண்டு-பாஸ் அணுகுமுறையைப் பயன்படுத்துகின்றன.

7) கருவி செயல்பாடு பொருத்துக.

அ. ஸ்கேன் செய்யும் போது ஆவணம் சரியாக ஒழுங்குபடுத்தப்படவில்லை எனில், உரையின் வரிகளை கிடைமட்டமாக அல்லது செங்குத்தாக மாற்ற சில டிகிரி கடிகார திசையில் அல்லது கடிகார திசையில் சாய்ந்து கொள்ள வேண்டியிருக்கும். - வரி நீக்கம் (Line removal)

ஆ. நேர்மறை மற்றும் எதிர்மறை புள்ளிகள், மென்மையான விளிம்புகளை அகற்றவும் - பைனரைசேஷன் (Binarisation)

இ. ஒரு படத்தை வண்ணம் அல்லது கிரேஸ்கேலில் இருந்து கருப்பு மற்றும் வெள்ளை நிறமாக மாற்றவும் (இரண்டு வண்ணங்கள் இருப்பதால் "பைனரி படம்" என்று அழைக்கப்படுகிறது) - டெஸ்பெகிள் (Despeckle).

ஈ. கிளிஃப் அல்லாத பெட்டிகளையும் கோடுகளையும் சுத்தம் செய்கிறது - டி-ஸ்கேவ் (De-skew)

8) கருவி செயல்பாடு பொருத்துக

அ. நெடுவரிசைகள், பத்திகள், தலைப்புகள் போன்றவற்றைத் தனித்துவமான தொகுதிகளாக அடையாளம் காட்டுகிறது – எழுத்து தனிமைப்படுத்தல் அல்லது "கூறுபடுத்தல்" (Character isolation or "segmentation")

ஆ. சொல் மற்றும் எழுத்து வடிவங்களுக்கான அடிப்படைகளை நிறுவுகிறது - எழுத்துவடிவத்தைப் புரிதல் (Script recognition)

இ. பன்மொழி ஆவணங்களில், எழுத்துவடிவம் சொற்களின் மட்டத்தில் மாறக்கூடும் - வரி மற்றும் சொல் கண்டறிதல் (Line and word detection)

ஈ.) ஒவ்வொரு எழுத்துக்குறி ஓசிஆருக்கு (per-character OCR), படக் கலைப்பொருட்கள் காரணமாக இணைக்கப்பட்டுள்ள பல எழுத்துக்கள் பிரிக்கப்பட வேண்டும் - தளவமைப்பு பகுப்பாய்வு அல்லது "ஸோனிங்" (Layout analysis or "zoning")

9) செயல்பாடு மென்பொருள் பொருத்துக.

அ. மூடிய கண்ணிகள்/வளையங்கள் (closed loops), வரி திசை (line direction) மற்றும் வரி குறுக்குவெட்டுகள் (line intersections) போன்ற "பண்புக்கூறுகளாக" கிளிஃப்களை சிதைக்கிறது - நவீன ஒளிமூலம் எழுத்துப் புரிதல் (OCR) மென்பொருள்

ஆ. உருப்படுத்தத்தின் பரிமாணத்தை குறைக்கிறது மற்றும் புரிதல் செயல்முறையைக் கணினி/கணக்கீட்டு அடிப்படையில் திறம்படச் செய்கிறது – கியூனிஃபார்ம் (Cuneiform) மற்றும் டெசராக்ட் (Tesseract) போன்ற மென்பொருள்கள்

இ. எழுத்துப் புரிதலுக்கு (character recognition) இரண்டு-பாஸ் அணுகுமுறையைப் பயன்படுத்துகின்றன - பிரித்தெடுக்கும் பண்புக்கூறுகள்

ஈ. ஓசிஆர்ஓபஸ் OCRopus அல்லது தெஸ்ஸெராக்ட் (Tesseract) போன்ற நரம்பியல் வலையமைப்புகளைப் (நெட்வொர்க்குகளைப்) பயன்படுத்துகிறது - பண்புக்கூறு பிரித்தெடுத்தல் கோடுகள் (lines)

10) பயன்பாடு- கருவி பொருத்துக

அ. தட்டச்சு செய்யப்பட்ட உரை, ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது – நுண்ணறிவு சொல் புரிதல் (IWR)

ஆ. தட்டச்சு செய்யப்பட்ட உரையை குறிவைக்கிறது, - நுண்ணறிவு எழுத்துப் புரிதல் (Intelligent word recognition (IWR))

இ. - கையால் எழுதப்பட்ட அச்சுஎழுத்து (handwritten printscript) அல்லது கர்சீவ் உரையை (cursive text) ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது - ஒளிவழி சொல் புரிதல் (Optical word recognition)

ஈ. கையால் எழுதப்பட்ட அச்சுஎழுத்து அல்லது கர்சீவ் உரையையும் குறிவைக்கிறது. - ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் (Optical character recognition (OCR))

3. சரியா/தவறா 15 வினாக்கள்

1) கணினியில் தரவை உள்ளிடுவதற்கான பாரம்பரிய வழி விசைப்பலகை வழியாகும்

√சரி/தவறு

2) எழுத்துணரி ஒரு தானியங்கா அடையாளம் காணும் கருவியாகும்.

சரி/தவறு √

3) அச்சிடப்பட்ட மற்றும் அச்சிடப்படாத எழுத்துக்கள் இரண்டும் எழுத்துணரியால் உணரப்படலாம்; ஆனால் செயல்திறன் நேரடியாக உள்ளீட்டு ஆவணங்களின் தரத்தைப் பொறுத்தது.

√சரி/தவறு

4) உள்ளீடு மிகவும் கட்டுப்படுத்தப்பட்டால், எழுத்துணரி அமைப்பின் செயல்திறன் சிறப்பாக இருக்காது.

சரி/தவறு√

5) எழுத்துணரி தொழில்நுட்பத்தின் முக்கிய நன்மைகள் நேரம் மிச்சப்படுத்தப்படுவது.

√சரி/தவறு

6) எழுத்துணரியின் ஏராளமான நன்மைகள் இருந்தாலும், இது முக்கியமாக வணிகத்தின் செயல்திறனை அதிகரிக்க வணிகங்களுக்கு உதவாது.

சரி/தவறு√

7) எழுத்துணரியைத் தேர்ந்தெடுப்பது, தரவு பிரித்தெடுப்பதை மேற்கொள்ள நிபுணர்களை பணியமர்த்துவதை குறைக்க வணிகங்களுக்கு உதவும்,

√சரி/தவறு

8) தரவு வகைப்பாட்டிற்கு எழுத்துணரியைப் பயன்படுத்தஇயலாது,

சரி/தவறு√

9) செயற்கை அறிவு அடிப்படையிலான எழுத்துணரி 2019 ஜப்பானிய கலாச்சாரம் மற்றும் செயற்கை அறிவு சிம்போசியத்தில் அறிமுகமானது.

√சரி/தவறு

10) எழுத்து உணர்தல் என்பது அமைப்பொழுங்கு உணர்தல் களத்தின் (pattern recognition area) துணைக்குழு அல்ல,

சரி/தவறு√

11) எழுத்து உணர்தலின் தோற்றம் உண்மையில் 1870ஆம் ஆண்டில் காணப்படுகிறது.

√சரி/தவறு

12) சி.ஆர். கேரி (C.R.Carey) விழித்திரை ஸ்கேனரைக் (retina scanner) கண்டுபிடித்த ஆண்டு 1860.

சரி/தவறு√

13) 1950களின் நடுப்பகுதியில் எழுத்துணரும் இயந்திரங்கள் வணிக ரீதியாக மாறி கிடைத்தது.

√சரி/தவறு

14) 1960 முதல் 1965 வரையிலான காலகட்டத்தில் தோன்றிய வணிக எழுத்துணரி ஒழுங்குமுறைகளை எழுத்துணரியின் இரண்டாம் தலைமுறை என்று அழைக்கலாம்.

சரி/தவறு√

15) இரண்டாம் தலைமுறையின் வாசிப்பு இயந்திரங்கள் 1960களின் நடுப்பகுதியிலும் 1970களின் முற்பகுதியிலும் தோன்றின.

√சரி/தவறு

4. ஒரு சொல்/சொற்றொடரில் விடைதருக (15 வினாக்கள்).

1) கணினியில் தரவை உள்ளிடுவதற்கான பாரம்பரிய வழி என்ன?

2) எழுத்துணரியின் செயல்திறன் எதை பொறுத்தது?

3) பகுதி எழுத்துக்களுக்கான செயற்கை அறிவு அடிப்படையிலான எழுத்துணரி எங்கு அறிமுகமானது?

4) எழுத்து உணர்தலின் தோற்றம் உண்மையில் எந்த ஆண்டில் காணப்படுகிறது?

5) விழித்திரை ஸ்கேனரைக் (retina scanner) கண்டுபிடித்த ஆண்டு எது?

6) விழித்திரை ஸ்கேனரைக் (retina scanner) கண்டுபிடித்தவர் யார்?

7) போலந்தைச் சார்ந்த பி. நிப்கோ என்ன கண்டுபித்தார்?

8) எழுத்துணரும் இயந்திரங்கள் வணிக ரீதியாக மாறி கிடைத்தது எப்போது?

9) முதல் உண்மையான எழுத்துணரி வாசிப்பு இயந்திரம் 1954இல் எங்கு நிறுவப்பட்டது?

10) 1960 முதல் 1965 வரையிலான காலகட்டத்தில் தோன்றிய வணிக எழுத்துணரி ஒழுங்குமுறைகளை எவ்வாறு அழைக்கலாம்?

11) இரண்டாம் தலைமுறையின் வாசிப்பு இயந்திரங்கள் எப்போது தோன்றின

12) மூன்றாம் தலைமுறை எழுத்துணரி எப்போது தோன்றியது?

13) மக்களுக்கான எழுத்துணரி எப்போது தோன்றியது?

14) ஒரு இயந்திரத்தை உருவாக்கி, எழுத்துக்களைப் படித்து அவற்றை நிலையான தந்தி குறியீடாக மாற்றியது யார்?

15)). ஒரே நேரத்தில், எட்மண்ட் ஃபோர்னியர் டி ஆல்பே (Edmund Fournier d'Albe) எதை உருவாக்கினார்?

5. ஒரு பத்தியில் விடை தருக (10 வினாக்கள்)

1) எழுத்துணர்தலின் வேறுபட்ட கட்டங்கள் குறித்து ஒரு பத்தியில் விடைதருக.

2) எழுத்துணையின் பயன்பாடுகளை பட்டியலிடுக.

3) எழுத்துணரிஅடிப்படையிலான தரவு உள்ளீட்டின் நன்மைகள் பற்றி ஒரு பத்தியில் விடைதருக.

4) எழுத்துணரி எவ்வாறு பயன்படுத்தப்படலாம்?

5) பார்வையற்ற மற்றும் பார்வை குறையுள்ள பயனர்களுக்கான எழுத்துணரி பற்றி சுருக்கமாக எழுதுக.

6) செயற்கை அறிவுடன் எழுத்துணரியை விரிவுபடுத்தல் பற்றி ஒரு பத்தி எழுதுக.

7) எழுத்துணரிகளின் வகைகளைப் பற்றி ஒரு பத்தியில் எழுதுக.

8) புத்திசாலித்தனமான எதிர்காலத்துடன் முதிர்ந்த தொழில்நுட்பம் பற்றி ஒரு பத்தி எழுதுக.

9) விலைப்பட்டியல் மற்றும் பல வகையான ஆன்வணங்களைச் செயலாக்குதல் பற்றி சுருக்கமாக எழுதுக.

10) பயன்பாடு சார்ந்த மேம்படுத்தல்கள் பற்றி சுருக்கமாக எழுதுக.

6.மூன்று பக்க அளவில் விடை தருக (5 வினாக்கள்)

1) எழுத்துணரியின் வரலாற்றை ஒரு கட்டுரையாக எழுத்துக.

2) எழுத்துணரியின் நுட்பங்கள் பற்று கட்டுரை வரைக.

3) எழுத்துணரியின் வேறுபட்ட பயன்பாடுகள் பற்றிய ஒரு கட்டுரை வரைக.

2. தமிழில் எழுத்துணரி

தமிழில் எழுத்துணரி (தமிழ் எழுத்துரு அறி மென்மம்) தொழில் நுட்பம்

ஒளிவழி எழுத்துணரி (Optical Character Recognition - OCR) என்பது எழுதப்பட்ட அல்லது அச்சடிகப்பட்ட எழுத்துக்களை கணினி உணரும்படி மாற்றும் ஒரு தொழில் நுட்பம் ஆகும். முதலில் ஆவணங்களை வருட வேண்டும் (scanning). இதற்கு நுட்பமான உணர்திறன் வாய்ந்த வருடி (scanner) தேவைப்படும். பின்னர் இதை மென்பொருள் மூலம் கணினிக்குப் புரியும்படி செய்யலாம். தமிழுக்கான ஒளி எழுத்துணரி உருவாக்கப்பட்டுப் பயன்பாட்டில் உள்ளது. ராமகிருஷ்ணன் மற்றும் பிறர் (Ramakrishnan 2000 Aparna and Ramakrishnan) தமிழில் ஒளி எழுத்துணரி உருவாக்குவது பற்றி விரிவாக விளக்குகின்றனர்.

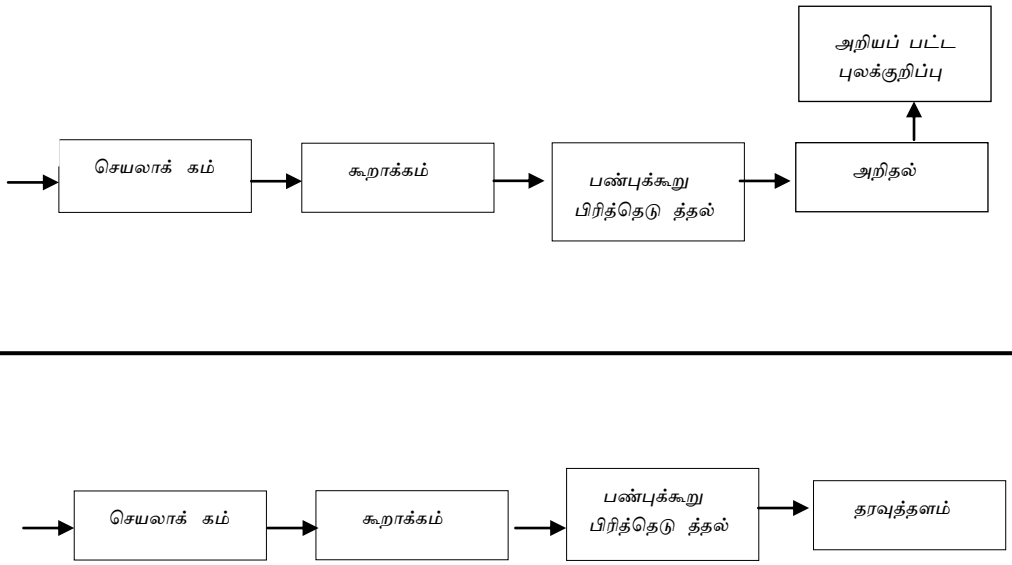
படச் செயலாக்கம் (Image processing) என்பது சில வழிமுறைகளால் படங்களின் தரத்தை மேம்படுத்தும் செயல்முறையாகும். படச் செயலாக்கம் என்பது சமிக்ஞை செயலாக்கத்தின் (signal processing) வடிவம் ஆகும். இங்கே உள்ளீடு ஒரு படம் மற்றும் படத்தின் வெளியீடு படத்துடன் தொடர்புடைய எழுத்துகளின் தொகுப்பாக இருக்கும். படச் செயலாக்க நுட்பத்தில், படம் இரு பரிமாணச் சமிக்ஞைகளாகக் கருதப்படும். படச் செயலாக்கம் என்பது ஒரு படத்தை மின்னிலக்க (டிஜிட்டல்) வடிவமாக மாற்றுவதற்கும், அதில் சில வேலைகளைச் செய்வதற்கும், மேம்பட்ட படத்தைப் பெறுவதற்கோ அல்லது அதிலிருந்து சில பயனுள்ள தகவல்களைப் பெறுவதற்கோ ஆகும். தமிழ் உரை கண்டறிதல் என்பது தட்டச்சு செய்யப்பட்ட அல்லது அச்சிடப்பட்ட உரையின் படங்களை இயந்திரக் குறியீட்டு உரையாக மாற்றும் இயந்திர அல்லது மின்னணு மாற்றமாகும். பாஸ்போர்ட் ஆவணங்கள், விலைப்பட்டியல், வங்கி அறிக்கை, ரசீதுகள், வணிக அட்டை, அஞ்சல் அல்லது பிற ஆவணங்கள் என அச்சிடப்பட்ட காகிதத் தரவு பதிவுகளிலிருந்து தரவு உள்ளீட்டு வடிவமாக இது பரவலாகப் பயன்படுத்தப்படுகிறது. இது அச்சிடப்பட்ட நூல்களை மின்னிலக்க (டிஜிட்டல்) மயமாக்குவதற்கான ஒரு பொதுவான முறையாகும், இதனால் இதை மின்னணு முறையில் திருத்தலாம், தேடலாம், மேலும் கச்சிதமாக சேமிக்கலாம், நிகழ்நிலையில் (ஆன்லைனில்) காட்டலாம், மேலும் இயந்திர மொழிபெயர்ப்பு, உரையிலிருந்து பேச்சு, முக்கிய தரவு (key data)

மற்றும் உரைச் சுரங்கம் (text mining) போன்ற இயந்திர செயல்முறைகளில் பயன்படுத்தலாம். தற்போது, கணினிகளில் தரவை உள்ளிடுவதற்கான பொதுவான வழியாக விசைப்பலகை உள்ளது. இது அநேகமாக அதிக நேரம் எடுக்கும் மற்றும் உழைப்பு மிகுந்த செயல்பாடாகும். ஒளிவழி எழுத்துணரி (OCR) என்பது மனித வாசிப்பின் இயந்திர பிரதி மற்றும் மூன்று தசாப்தங்களுக்கும் மேலாக தீவிர ஆராய்ச்சிக்கு உட்பட்டது. வருடப்பட்ட அல்லது ஸ்கேன் செய்யப்பட்ட படங்களின் எந்திர/மெக்கானிக்கல் அல்லது மின்/எலக்ட்ரானிக் மாற்றமாக எழுத்துணரியை (OCR) விவரிக்கலாம். படங்களைக் கையால் எழுதலாம், தட்டச்சு செய்யலாம் அல்லது அச்சிடலாம். எழுத்துணரி இவற்றை மின்னிலக (டிஜிட்டல்) மயமாக்கும் ஒரு முறையாகும்; இதனால் அவை மின்னணு முறையில் தேடப்பட்டு இயந்திர செயல்முறைகளில் பயன்படுத்தப்படலாம்.

ஒளிவழி எழுத்துணர்தல் (Optical Character Recognition (OCR)), பொதுவாக ஓசிஆர் (OCR) என சுருக்கமாக அழைக்கப்படுகிறது, இது கையால் எழுதப்பட்ட, தட்டச்சு செய்யப்பட்ட அல்லது அச்சிடப்பட்ட உரையின் வருடப்பட்ட (ஸ்கேன் செய்யப்பட்ட) படங்களின் இயந்திர அல்லது மின்னணு மொழிபெயர்ப்பாகும். இந்தப் பயன்பாட்டில் பயனர் ஒரு படத்தைத் தனது ஸ்மார்ட் தொலைபேசியில் வருடப்பட்ட படமாக இடுவார். வருடப்பட்ட படங்கள் டி-ஸ்கூவிங்கில் (de-skewing) ஈடுபடும். படம் டி-ஸ்கூ (Image De-skew) என்பது படங்களிலிருந்து வளைவை அகற்றும் செயல்முறையாகும் (குறிப்பாக வருடியைப் பயன்படுத்தி உருவாக்கப்பட்ட பிட்மேப்கள்). ஸ்கூ (Skew) என்பது வருடப்பட்ட படங்களில் ஏற்படக்கூடிய ஒரு நிலை. புகைப்படக்கருவி தவறாக வடிவமைக்கப்பட்டிருக்கலாம்; மேற்பரப்பில் குறைபாடுகள் இருக்கலாம்; வருடும்போது காகிதம் முற்றிலும் தட்டையாக வைக்கப்படாதிருக்கலாம். பிரித்தெடுக்கப்பட்ட படிகளைத் தொடர்ந்து கைப்பற்றப்பட்ட படத்தின் இருமையாக்கம், எழுத்துப் பிரிப்பு, மற்றும் நோக்குநிலை என்பன வரும். இந்த படிகளைக் கடந்து உரை படத்திலிருந்து பிரித்தெடுக்கப்படும். படத்திலிருந்து உரை பிரித்தெடுக்கப்பட்டதும், எடுத்துக்கொண்ட நோக்கத்திற்காகப் பதிவேற்றப்படும்

ஆவணப் படச் செயலாக்கம் (Document Image processing) மற்றும் ஒளி எழுத்துணரி (ஆப்டிகல் கேரக்டர் ரெகக்னிஷன்) (Optical Character Recognition (OCR) ஆகியவை கடந்த சில பதின்ம ஆண்டுகளாக மனித-இயந்திர இடைமுகத் (human-machine interface) துறையில் முன்னணி ஆராய்ச்சிப் பகுதியாகும். அறிவார்ந்த வருடல் இயந்திரங்கள், உரையில் இருந்து பேச்சு மாற்றிகள் (text to speech converters) மற்றும் தானியங்கி மொழியிலிருந்து மொழி மொழிபெயர்ப்பிகள் (automatic language to language translators) போன்ற பயன்பாடுகளில் இயந்திரம் அச்சிடப்பட்ட அல்லது கையால் அச்சிடப்பட்ட ஆவணத்தை அறிவது/புரிவது ஒரு முக்கிய பகுதியாகும். ஆவணப் படப் பகுப்பாய்வின் (document image analysis) நோக்கம் காகித ஆவணத்தில் உள்ள உரை மற்றும் வரைகலைக் (கிராபிக்ஸ்) கூறுகளை அறிவதும்/புரிவதும், மனிதர்களைப் போலவே நோக்கம் கொண்ட தகவல்களைப் பெறுவதும் ஆகும். ஆவண படப் பகுப்பாய்வின் இரண்டு கூறுகள் உரைச் செயலாக்கம் (Textual processing) மற்றும் வரைகலை செயலாக்கம் (Graphical processing) என்பன ஆகும். உரைச் செயலாக்கம் ஆவண படத்தின் உரை கூறுகளைக் கையாள்கிறது. வரைகலைச் செயலாக்கமானது வரி வரைபடங்களை உருவாக்கும் உரை அல்லாத வரி மற்றும் குறியீட்டுக் கூறுகளைக் கையாளுகிறது, உரைப் பிரிவுகள் மற்றும் நிறுவனத்தின் சின்னங்களுக்கு (company logos) இடையில் நேர் கோடுகளை வரையறுக்கிறது. தற்போதைய சூழலில், நாம் உரைச் செயலாக்கப் பகுதிக்கு மட்டுப்படுத்துகிறோம்.

ஒளி எழுத்துணரி அமைப்பின் தொகுதி வரைபடம் (படம் 1)



படம் 1 இல் உள்ள தொகுதி வரைபடத்தில் காட்டப்பட்டுள்ளபடி, ஒரு எழுத்துணரி (OCR) ஒழுங்குமுறை இரண்டு கட்டங்களைக் கொண்டுள்ளது, அதாவது பயிற்சி கட்டம் மற்றும் உணர்தல் கட்டம். இரண்டு கட்டங்களும் சில பொதுவான படிகளைக் கொண்டுள்ளன. வளைவு திருத்தம் (skew correction) என்பது பைனரி ஆவணப் படத்தில் மேற்கொள்ளப்படும் முதல் செயல்பாடாகும். ஸ்கேனர் படுக்கையில் ஒரு காகிதத்தை வைக்கும்போது சாய்வது தவிர்க்க முடியாதது. இந்த சாய்வு கோணம் பொதுவாக ஆவணத்தின் வளைவு கோணம் (skew angle) என்று அழைக்கப்படுகிறது. உரைக் கோடுகளைக் கிடைமட்டமாக்குவதற்கு வளைவு கோணத்தைக் கண்டறிந்த பின் படத்தை மீண்டும் சுழற்ற வேண்டும். அடுத்து, வளைவு சரிசெய்யப்பட்ட படத்தில் ஆவணப் பிரிவு செய்யப்படுகிறது. ஆவணப் பிரிவில், உரை கோடுகள் முதலில் பிரிக்கப்படுகின்றன. பிரிக்கப்பட்ட ஒவ்வொரு உரை வரிக்கும், அதில் உள்ள சொற்கள் பிரிக்கப்படுகின்றன. ஒவ்வொரு வார்த்தைக்கும், தனிப்பட்ட சின்னங்கள் பிரிக்கப்பட்டு, அறிபவருக்கு (recognizer) உள்ளீடாக வழங்கப்படுகின்றன. அறியப்பட்ட அனைத்துச் சின்னங்களையும் நிலையில் வைத்து, அறியப்பட்ட சொல் குறியாக்கம் செய்யப்படுகிறது. அதே வழியில், உரை வரியைக் குறியாக்க வார்த்தைகள்

ஒன்றாக இணைக்கப்படுகின்றன. குறியீடு இயல்பாக்குதல் (Symbol normalization) என்பது குறியீட்டு அறிதலின் முதல் படியாகும். பிரிக்கப்பட்ட ஒவ்வொரு சின்னமும் இயல்பாக்கப்பட்ட அளவுக்குக் கொண்டு வரப்பட்டு குறியீட்டின் உருகோடு (skeleton of the symbol) கண்டறியப்பட்டது. அறிதல் செயல்முறையை எழுத்துரு மற்றும் அளவிலிருந்து சுதந்திரமாக்குவதற்கான முக்கிய படியாகும். தனிப்பட்ட குறியீடுகளை இயல்பாக்கப்பட்ட அளவிற்கு கொண்டு வர குறியீட்டு இயல்பாக்கம் அவசியம், இதனால் அவை குறிப்பு தரவுத்தளத்தில் அறியப்பட்ட குறியீடுகளுடன் ஒப்பிடலாம். ஒவ்வொரு இயல்பாக்கப்பட்ட குறியீட்டிலிருந்தும், தொடர்புடைய அம்சங்கள் பிரித்தெடுக்கப்படுகின்றன. அம்சங்கள் என்பது குறியீட்டின் வடிவ தகவல்களைச் சுமக்கும் திசையன்களின் தொகுப்பாகும். நடைமுறையில் ஏராளமான அம்சம் பிரித்தெடுக்கும் நடைமுறைகள் உள்ளன. பிரித்தெடுக்கப்பட்ட அம்சங்களின் அடிப்படையில், எழுத்துக்கள் வெவ்வேறு வகுப்புகளாக வகைப்படுத்தப்படுகின்றன. வகைப்படுத்தி அறியப்படாத குறிட்யீட்டின் பண்புக்கூறு திசையனைத் (feature vector), தரவுத்தளத்தில் அறியப்பட்ட குறியீடுகளுடன், சில தூர அளவின் அடிப்படையில் ஒப்பிட்டு, அந்த குறியீட்டை முன் வரையறுக்கப்பட்ட வகைப்பாடு விதிப்படி வகைப்படுத்துகிறது.

தமிழ் எழுத்துக்களின் பண்புகள்

தமிழ் எழுத்து முறையில் 12 உயிரெழுத்துக்கள் மற்றும் 23 மெய் எழுத்துக்கள் உள்ளன. மற்ற அனைத்து இந்திய மொழிகளையும் போலவே தமிழ் எழுத்து முறையின் அடிப்படை அமைப்பும் ரோமன் எழுத்து முறையிலிருந்து வேறுபட்டது. ஒலிப்பு ரீதியாக, உயிரெழுத்துக்கள் தனியாக நிகழலாம் அல்லது மெய் அல்லது மற்றொரு உயிரெழுத்துடன் இருக்கலாம். உயிரெழுத்துடன் இல்லாத உயிரெழுத்து மற்றும் சில அடிக்கடி பயன்படுத்தப்படும் உயிரெழுத்துக்கள் தனித்தனி குறியீட்டு பிரதிநிதித்துவங்களைக் கொண்டுள்ளன. மெய்யெழுத்துடன் சேரும்போது, அதனுடன் இணைந்த மெய்யெழுத்துக்கு 'மாத்திரை' சேர்க்கும் வடிவத்தில் குறியீட்டாக்கம் உள்ளது. தனியாக மெய் தோன்ற முடியாது. ஒரு மெய்யின் முன் அல்லது பின் அல்லது முன்னும் பின்னும் ஒரு உயிரெழுத்து இருக்க வேண்டும். மெய்யெழுத்துக்கு முந்தைய உயிரெழுத்து, வார்த்தையின் முந்தைய எழுத்தின் ஒரு பகுதியாகக் கருதப்படுகிறது. ஒரு உயிரெழுத்து ஒரு வார்த்தையில் ஒரு

மெய்யைப் பின்தொடரும்போது, சம்பந்தப்பட்ட உயிரெழுத்துடன் தொடர்புடைய 'மாத்திரை' மெய்யின் அடையாளத்தை மாற்றியமைக்கிறது. /அ/-க்கு எந்த மாத்திரையும் இல்லை, இது எழுத்துக்களின் தொகுப்பில் முதல் உயிரெழுத்து ஆகும். /அ/ ஒரு மெய்யைப் பின்தொடர்ந்தால், மெய் அதன் அசல் வடிவத்தில் தோன்றும் (எ.கா. க்+அ=க). எந்த உயிரெழுத்து மெய்யையும் பின்பற்றவில்லை என்றால், அசல் மெய்யின் மேல் ஒரு புள்ளியைச் ("்") சேர்ப்பதன் மூலம் இது குறிக்கப்படுகிறது.

தேவநாகரி அல்லது பெங்காலி போன்ற பல இந்திய எழுத்துமுறைகளில், நிறுத்த மெய்யெழுத்தைத் தொடர்ந்து வரும் மெய் சில நேரங்களில் கூட்டு மெய் என எழுதப்படுகிறது, இது ஒரு அடையாளமாகும். தமிழில் அத்தகைய சின்னம் இல்லை. இருப்பினும் தமிழில் ஒரு மெய்யை ஒரு மாத்திரையுடன் சேர்க்கும்போது மற்ற இந்திய மொழிகளைப் போலல்லாமல் முழுமையாக மாற்றியமைக்கப்படலாம். தமிழ் விஷயத்தில் தலைப்பு அல்லது 'ஷிரோரேகா' (shirorekha) இல்லை. எனவே எழுத்துக்கள் ஒன்றுக்கொன்று பிரிக்கப்படுகின்றன. வெவ்வேறு மெய் மாற்றும் போது அதே உயிரெழுத்து வெவ்வேறு வழிகளில் மாற்றப்படலாம்; அதாவது கொடுக்கப்பட்ட உயிரெழுத்துக்கான 'மாத்திரை'யின் வடிவம் தனித்துவமாக இருக்காது. 'மாத்திரைகள்' /உ/ மற்றும் / ஊ/-க்கு இது பொதுவான வழக்கு. இது பிற பல இந்திய மொழிகளுக்குப் பொதுவானதல்ல. படம் 2 தமிழ் மொழியின் அடிப்படை எழுத்துக்குறிகளைக் காட்டுகிறது.

தமிழ் மொழியின் அடிப்படை எழுத்துத் தொகுப்பு (படம் 2)

தமிழ் உயிரெழுத்துக்கள்	அ, ஆ, இ, ஈ, உ, ஊ, எ, ஏ, ஐ, ஒ, ஓ, ஔ
தமிழ் மெய்யெழுத்துக்கள்	க, ச, ட, த, ப, ற, ஞ, ங, ண, ம, ந, ய, ர, ல, வ, ழ, ள
பயன்பாட்டில் உள்ள வடமொழி எழுத்துக்கள்	ஜ, ஸ, ஷ, ஹ, க்ஷ,

- **அனுமானங்கள்**

பின்வருவன எழுதுணரியின் திறமையான செயல்பாட்டுக்காகச் செய்யப்பட்ட சில அனுமானங்கள்:

- ஆவணத்தில் படங்கள் இல்லாத உரை மட்டுமே இருக்க வேண்டும்.
- ஒரு பக்கத்தில் உள்ள நெடுவரிசைப் பிரிவு கருதப்படவில்லை.
- அச்சிடப்பட்ட உரை ஆவணம் மட்டுமே கருதப்படுகிறது.
- எழுதுணரி தமிழில் மட்டுமே உள்ள ஆவணங்களுக்கு நல்ல அறிதலை அளிக்கிறது.

முன் செயலாக்கம்

வளைவு கண்டறிதல்

ஒரு ஆவணம் ஒளிவழி வருடிக்குக் (ஆப்டிகல் ஸ்கேனர்) கைமுறையாக வழங்கப்படும்போது, சில டிகிரி வளைவு (சாய்வு) தவிர்க்க முடியாதது. வளைவு கோணம் என்பது உரை கோடுகள் கிடைமட்ட திசையுடன் செய்யும் கோணம். வளைவு மதிப்பீடு மற்றும் திருத்தம் உரை செயலாக்கத்தில் மிக முக்கியமான படிகளில் ஒன்றாகும். ஒருங்கிணைந்த உரை மற்றும் கிராபிக்ஸ் சூழ்நிலையில் கூட, உரை பகுதியிலிருந்து வளைவு மதிப்பீடு செய்யப்படுகிறது, ஏனெனில் உரை பகுதி வளைவு மதிப்பீட்டிற்கு சாதகமான கட்டமைப்பைக் கொண்டுள்ளது. ஒரு துல்லியமான வளைவு கண்டறிதல் முறை இங்கு முன்மொழியப்படுகிறது; இது ஒரு ஆவணத்தின் வளைவை இரண்டு படிகளில் கண்டறிகிறது.

கரடுமுரடான வளைவு மதிப்பீடு (Coarse skew estimation)

கரடுமுரடான வளைவு மதிப்பீட்டில் சம்பந்தப்பட்ட படிகள் பின்வருமாறு.

- ரன்-நீளம் சீர்படுத்தல் (Run-length Smoothing)
- ”இடைக்கால வரி கண்டறிதல்” (Interim Line detection) மற்றும் “இடைக்கால வரி படத்தை” (Interim Line Image) உருவாக்குதல்.
- கரடுமுரடான வளைவுக் கோணத்தை மதிப்பிடுவதற்கு "இடைக்கால வரி படத்தில்" ஹஃப் டிரான்ஸ்ஃபார்ம் (Hough Transform) பயன்பாடு.

ஓட்ட-நீளம் (run-lenth) சீர்படுத்துதல் என்பது எழுத்துக்களுக்குள்ளும் எழுத்துக்களுக்குள்ளும் உள்ள இடைவெளிகளை நிரப்புகிறது. இதனால் ஒவ்வொரு வார்த்தையும் இணைக்கப்பட்ட கூறுகளாக மாறும். ஓட்டத்தின் நீளத்தைச் சீர்படுத்துவதற்கு ஒரு தொடக்க நிலையக்காட்டிலும் குறைவான நீளத்தைக் கொண்ட அனைத்து பின்னணி ஓட்ட-நீளங்களும் முன்னணி ஓட்டநீளங்களாக மாற்றப்படுகின்றன.

அடுத்த கட்டம் இடைக்கால கோடுகளைக் கண்டுபிடிப்பதாகும். இடைக்கால வரி என்பது இரண்டு உரை வரிகளுக்கு இடையிலான பின்னணி இடைவெளியை மத்தியஸ்தம் செய்யும் வரி. எக்ஸ்-அச்சின் நேர்மறையான திசையைப் பொறுத்து இடைக்காலக் கோடுகளின் நோக்குநிலை படத்தின் வளைவு கோணத்திற்கு சமம். அடுத்து, வளைவு கோணத்தின் தோராயமான மதிப்பீடு x-அச்சின் நேர்மறையான திசையைப் பொறுத்து இடைக்காலக் கோடுகளின் நோக்குநிலை கோணமாகப் பெறப்படுகிறது. ஹஃப் டிரான்ஸ்ஃபார்ம் இங்கே பயன்படுத்தப்படலாம். பயன்படுத்தப்படும் ஒரு நேர் கோட்டின் சமன்பாட்டின் வடிவம்: $y \cos \theta - x \sin \theta = P$, இங்கு P என்பது தோற்றத்திலிருந்து கோட்டின் செங்குத்தாக உள்ள தூரம் மற்றும் x என்பது கோட்டிற்கும் x அச்சின் நேர்மறை திசைக்கும் இடையிலான கோணம் ஆகும். பெறப்பட்ட முடிவுகளிலிருந்து, இந்த முறை உண்மையான மதிப்பைப் பற்றி ± 0.25 இன் வரம்பிற்குள் வளைவு கோணத்தைக் கண்டறிய உதவுகிறது.

நன்றாக வளைவு கண்டறிதல் (Fine skew detection)

நன்றாக வளைவு (Fine skew) கண்டறிதல் பின்வரும் படிகளை உள்ளடக்கியது:

- ஒரு சிதறல் படத்தை உருவாக்கப் பிரிக்கப்பட்ட வரி படங்களின் மேற்பொருத்தம் (சூப்பர்போசிஷன்/Superposition).
- சிதறல் படத்தின் முதன்மை அச்சுகளைக் கண்டறிதல்.

கரடுமுரடான மதிப்பீட்டிலிருந்து வளைவு கோணத்தின் அறிவைப் பயன்படுத்தி, உரை கோடுகள் பிரிக்கப்படுகின்றன. அசல் படத்தின் திட்ட சுயவிவரத்தை வளைவு கோணத்தின் கரடுமுரடான மதிப்பீட்டிற்கு சமமான கோணத்தில் கண்டுபிடிப்பதன் மூலம் உரை வரி பிரிவு செய்யப்படுகிறது. அனைத்து உரை வரிகளும் இந்த வழியில் பிரிக்கப்பட்டு, அவற்றின் மையங்கள் ஒரே நேரத்தில் இருக்கும் வகையில் ஒன்றுக்கொன்று மேற்பொருத்தம் செய்யப்படுகின்றன.

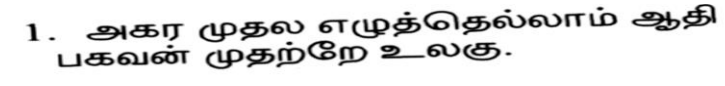
அவ்வாறு உருவான படத்தை சிதறல் படம் (scatter image) என்று அழைக்கப்படுகிறது. இதன் விளைவாக வரும் சிதறல் படத்தின் முதன்மை அச்சின் திசையானது நேர்த்தியான வளைவு திசையாக எடுக்கப்படுகிறது. பெறப்பட்ட முடிவுகள் இறுதி மதிப்பீட்டின் துல்லியம் +/- 0.060 என்பதைக் காட்டுகிறது.

வளைவுத் திருத்தம் (Skew Correction)

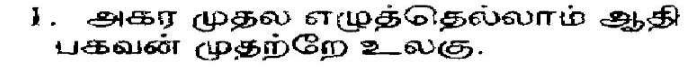
வளைவுப் பிழை கண்டறியப்பட்டதும் வளைவு திருத்தப்பட வேண்டும். வளைவு திருத்தும் இந்த செயல்முறை படத்தை வளைவுக்குச் சமமான கோணத்தில் எதிர்த் திசையில் சுழற்றுவதன் மூலம் அடையப்படுகிறது. வளைவு திருத்தும் செயல்முறை பொதுவாக இருமையாகப்பட்ட (binarised) படத்தில் செய்யப்படுகிறது. பயன்படுத்தப்படும் சுழற்சி வழிமுறை பிலினியர் இடைக்கணிப்பைப் (bilinear interpolation) பயன்படுத்துகிறது. ஆனால் அளவீட்டு விளைவுகள் (quantisation effects) காரணமாக வளைவு சரிசெய்யப்பட்ட விளைவுப் படம் நிறைய சிதைவுகளைக் கொண்டுள்ளது. எனவே இருமைப் படத்தை (binary image) விட சாம்பல் அளவிலான படத்தில் (gray scale image) வளைவுத் திருத்தம் செய்வதன் மூலம் அளவீட்டு விளைவுகளைக் குறைக்க இங்கு முயற்சி மேற்கொள்ளப்பட்டுள்ளது. வெளியீடுகள் கீழே உள்ள படம் 3 இல் காட்டப்பட்டுள்ளன.

படம் 3: வளைவு திருத்தத்தின் வெளியீடுகள்

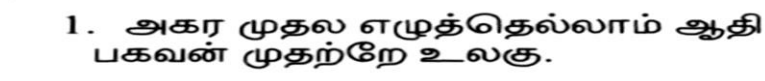
படம் 3அ: அசல் வளைந்த படம்



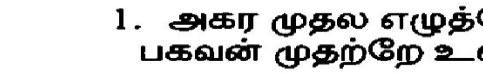
படம்3ஆ: இருமைப் படத்தில் வளைவு திருத்தம் செய்யப்பட்டது



படம் 3 இ: வளைவு சரி செய்யப்பட்ட சாம்பல் அளவிலான படம்

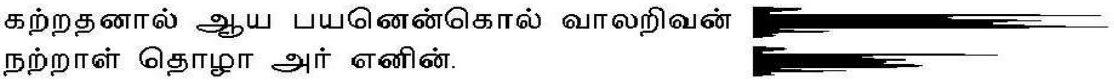


படம் 3 ஈ: வளைவு சரி செய்யப்பட்ட இருமையாக்கப்பட்ட படம்



கூறிடல் (segmentation) என்பது ஒரு படத்திலிருந்து ஆர்வமுள்ள பொருட்களைப் பிரித்தெடுக்கும் செயல்முறையாகும். ஆவணப் பகுப்பாய்வில், உரை அறிதலை நோக்கிய முதல் செயல்பாடு உரைப் பகுதிகளை கிராபிக்ஸ், வரைபடங்கள் மற்றும் பிற புள்ளிவிவரங்களிலிருந்து பிரிப்பதாகும். பிரிவின் முதல் படி கோடுகளைக் கண்டறிதல். பின்னர் வரியில் உள்ள வார்த்தையைக் கண்டறிந்து, அந்த வார்த்தையின் தனிப்பட்ட தன்மையைக் கண்டறியும். கிடைமட்ட முந்திட்ட தோற்றவடிவத்தின் உதவியுடன் உரை வரிகளை அடையாளம் காணலாம். ஒரு குறிப்பிட்ட திசையில் ஒரு ஆவணத்தின் முந்திட்ட தோற்றவடிம் அந்த திசையில் பிக்சல்களின் இயங்கும் தொகை ஆகும். தோற்றவடிவம் வெற்றிடப் புள்ளிகளை வரி எல்லைகளில் காட்சிப்படுத்துகிறது மற்றும் இந்த குறைந்தபட்சப் புள்ளிகளின் இருப்பிடம் வரி எல்லைகளைக் குறிக்கிறது. இருமைப் படங்களைப் (binary images) பொறுத்தவரை, சுயவிவரம் பூஜ்ஜியத்திற்குச் செல்லும் புள்ளிகள் இவை. கீழேயுள்ள படம் 4 இருமையாக்கம் செய்யப்பட்ட படத்தின் கிடைமட்ட முந்திட்ட தோற்றவடிவத்தைக் (horizontal projection profile) காட்டுகிறது.

படம் 4: கிடைமட்ட முந்திட்ட தோற்றவடிவங்களுடன் உரை கோடுகள்



செங்குத்து முந்திட்ட தோற்றவடிவத்தின் உதவியுடன் சொல் கூறு செய்யப்படலாம். செங்குத்து முந்திட்ட தோற்றவடிவம் சொல் இடைவெளிகளுடன் தொடர்புடைய புள்ளிகளில் வெற்றிடங்களைக் காட்டுகிறது. இந்த சிறுமப் புள்ளிகளின் உதவியுடன் இந்த சொல் எல்லைகளை அடையாளம் காணலாம். கீழேயுள்ள படம் 5 இருமையாக்கப்பட்ட படத்தின் செங்குத்து முந்திட்ட தோற்றவடிவத்தைக் காட்டுகிறது.

படம் 5: தொடர்புடைய செங்குத்துத் திட்ட சுயவிவரங்களுடன் உரை கோடுகள்

எழுத்துக் கூறிடல் (character segmentation) என்பது சொற்களிலிருந்து தனிப்பட்ட சின்னங்களை/குறியீடுகளைப் பிரிக்கும் செயல்முறையாகும். இந்தத் தனிப்பட்ட கூறுகள் பொதுவாக அவற்றுக்கிடையே செங்குத்து இடைவெளியைக் கொண்டுள்ளன. தமிழ், அதன் அடிப்படை கட்டமைப்பின் காரணமாக, மேல் மற்றும் கீழ் இணைக்கப்பட்ட அல்லது அடிப்படை எழுத்துக்களிலிருந்து வேறுபட்ட மாற்றிகளைக் கொண்டுள்ளது. அத்தகைய சூழ்நிலையில், ஒரு எளிய கூறிடல் திட்டத்தைப் பயன்படுத்தலாம். இருப்பினும், சாய்வு எழுத்துருக்களுக்கான எழுத்துக்களுக்கு இடையில் பூஜ்ஜியங்களின் செங்குத்து இயக்கம் இல்லாத நிலையில், சுயவிவர அடிப்படையிலான அணுகுமுறையை மேற்கொள்ள முடியாது. எனவே, தனிப்பட்ட சின்னங்களை பிரிக்க, இணைக்கப்பட்ட கூறு அணுகுமுறை (connected component approach) பயன்படுத்தப்படுகிறது. இந்த நுட்பத்தில், அனைத்து முன்புற/முன்னணி பிக்சல்களும் அண்டை பிக்சல்களுடனான இணைப்புக்காகச் சரிபார்க்கப்பட்டு சரியான முறையில் பெயரிடப்பட்டுள்ளன. பூஜ்ஜியமற்ற அண்டை என்று பெயரிடப்பட்ட பிக்சல்களுக்கு அண்டைகளின் லேபிள்/புலக்குறிப்பு வழங்கப்படுகிறது மற்றும் இல்லாதவைகளுக்குப் புதிய லேபிள்/புலக்குறிப்பு ஒதுக்கப்படுகிறது. ஒரே லேபிள்/புலக்குறிப்பு மதிப்புகளைக் கொண்ட பிக்சல்களைத் தனிமைப்படுத்துவதன் மூலம் இறுதிக் கூறிடல் செய்யப்படுகிறது. இந்த முறை எந்த எழுத்துருடனும் சரியாக வேலை செய்கிறது.

சின்னம் முன் செயலாக்கம் (Symbol Pre-processing)

எந்தவொரு எழுத்துணரி (OCR) வடிவமைப்பிலும் எழுத்து முன் செயலாக்கம் ஒரு முக்கிய பகுதியாகும். எழுத்து முன் செயலாக்கத்தைப் போல வேறு எந்த ஆவணச் செயலாக்க நடவடிக்கையும் முக்கியமில்லை. அசல் உரை படத்திலிருந்து பிரிக்கப்பட்ட/கூறாக்கப்பட்ட எழுத்துக்கள் வெவ்வேறு அளவுகளில் உள்ளன. ஒரு குறிப்பிட்ட பண்புக்கூறு இடைவெளியில் (particular feature space) ஒப்பிடுவதற்கு, ஒரே குழுவைச்/கிளஸ்டரைச் சேர்ந்த எழுத்துக்கள் இயல்பாக்கப்பட்ட அளவுக்குக் கொண்டு வரப்பட வேண்டும். எனவே, எழுத்துக்களை இயல்பாக்கப்பட்ட அளவிற்கு அளவிடுதல் ஒரு முக்கியமான படியாகும். மெல்லியதாக இருப்பது ஒரு விருப்பமான படியாகும், பொதுவாக உணரி (recognizer) எழுத்தின் திண்மையிலிருந்து சுதந்திரமாக வடிவமைக்கப்பட்டிருந்தால் அது தவிர்க்க முடியாதது.

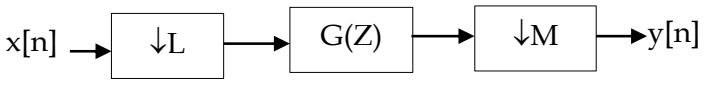
ஒரு படத்தை அளவிடுவது இரண்டு செயல்பாடுகளின் கலவையாகக் கருதப்படலாம். முதலில், ஒவ்வொரு வரிசையையும் விரும்பிய அகலத்திற்கு அளவிட முடியும், பின்னர், வரிசை-இயல்பாக்கப்பட்ட படத்தின் ஒவ்வொரு நெடுவரிசைகளும் தேவையான படத்தைப் பெற விரும்பிய உயரத்திற்கு அளவிடப்படலாம். எனவே, படம் இரு பரிமாண சமிக்ஞையாக இருந்தாலும், இயல்பாக்கம் செயல்முறை என்பது பல பரிமாண மறுசீரமைப்பு செயல்முறைகளின் கலவையாகும். படம் ஒரு குறிப்பிட்ட திசையில் நீட்டப்பட்டால், அது மேம்பாடு மற்றும் குறைத்தல் என்பது சுருங்குவதைக் குறிக்கிறது.

பட மறு மாதிரியின் ஒரு பிரபலமான முறை இருநேரியல் இடைச்செருகல் (bilinear interpolation) ஆகும். இருநேரியல் இடைச்செருகலின் குறைபாடு படத்தின் மெல்லிய பகுதிகளில் தகவல்களை இழப்பதாகும். இந்த குறைபாட்டைச் சமாளிக்க, மல்டிரேட் சிக்னல் செயலாக்கத்தின் பார்வையில் இருந்து சிக்கலை அணுகுவோம். நீளம் Mk இன் சமிக்ஞையை ஒரு நீள Lk க்கு மீண்டும் மாதிரி செய்ய (k என்பது Mk மற்றும் Lk இன் மிகப் பெரிய பொதுவான காரணியாகும்), நாம் ஒரு விரிவாக்கி வழியாகச் சமிக்ஞையை அனுப்புகிறோம், அதன்பிறகு குறைந்த பாஸ் வடிப்பான் மற்றும் ஒரு டெசிமேட்டர். விரிவாக்கியின் விரிவாக்க விகிதம் L, டெசிமேட்டரின் டெசிமேஷன் காரணி M மற்றும் குறைந்த பாஸ் வடிப்பானின் பாஸ்பேண்ட் -π / l முதல் + π / l வரை இருக்கும், அங்கு l = அதிகபட்சம் (L, M). கணினியின் தொகுதி வரைபடம் படம் 6 இல் காட்டப்பட்டுள்ளது. எல்.பி.எஃப் (LPF) மற்றும் டெசிமேட்டரின் கலவையானது கணக்கீட்டைக் குறைக்க பாலிஃபேஸ் உணர்தலாக (polyphase realization) செயல்படுத்தப்படலாம். காரணமற்ற குறைந்த பாஸ் வடிப்பான்களை உணர முடியும், ஏனெனில் காரணக் கண்ணோட்டத்தில் எந்த தடையும் இல்லை.

அதிர்வெண் களத்தில், மறுவடிவமைப்பு செயல்முறை M > L (M < L) வழக்கில் M / L இன் காரணி மூலம் ஸ்பெக்ட்ரமின் நீட்சிக்கு (சுருங்கி) சமம். L > M வழக்கில், சமிக்ஞையின் டிஎஃப்டி திசையனின் Mk/2 வது குணகத்தைச் சுற்றிச் சமச்சீராகத் திணித்தல் (L-M)k பூஜ்ஜியங்களால் மறுவடிவமைப்பு அடைய முடியும். L < M என்றால், சமிக்ஞையின் டிஎஃப்டி திசையனின் எம்.கே /

2-வது குணகத்தைச் சுற்றி (M-L) கே எண் குணகங்களை சமச்சீராகக் குறைக்க முடியும். இதன் விளைவாக துண்டிக்கப்பட்ட அல்லது பூஜ்ஜியம் திணிக்கப்பட்ட DFT இன் ஐடிஎஃப்டி (IDFT) மறு மாதிரி சமிக்ஞையாகும். குறைவான மாதிரியின் விஷயத்தில், டி.எஃப்.டி.யின் பூஜியமல்லாத்த (நொன்ஜெரோ) குணகங்களைக் குறைப்பது சமிக்ஞையின் மாற்றுப்பெயர்வைத் தவிர்ப்பதற்காக குறைந்த-பாஸ் வடிகட்டலுடன் ஒத்திருக்கிறது மற்றும் அதனுடன் தொடர்புடைய தகவல்களை இழக்கிறது.

படம் 6: மறு மாதிரி செயலாக்கத்தின் தடுப்பு வரைபடம். x [n] என்பது மீண்டும் மாற்றப்பட வேண்டிய சமிக்ஞையாகும். y [n] என்பது மறுசீரமைக்கப்பட்ட சமிக்ஞையாகும்.



மெல்லியதாக இருப்பது வட்டி பொருளை ஒரு பிக்சல் அகலத்தின் விளிம்பாக மாற்றும் செயல்முறையாகும். இங்கே [2] இல் கொடுக்கப்பட்ட மெல்லிய வழிமுறை பயன்படுத்தப்பட்டுள்ளது. கீழே உள்ள படம் 7 ஒரு கூறிடப்பட்ட எழுத்து மற்றும் அதனுடன் இயல்பாக்கப்பட்ட மற்றும் மெல்லிய படங்களைக் காட்டுகிறது.

படம் 7: இயல்பாக்கப்பட்ட மற்றும் மெல்லிய படங்கள்

படம் 7 அ: அசல் அளவு 28 x 28-இன் கூறிடப்பட்ட சின்னம்

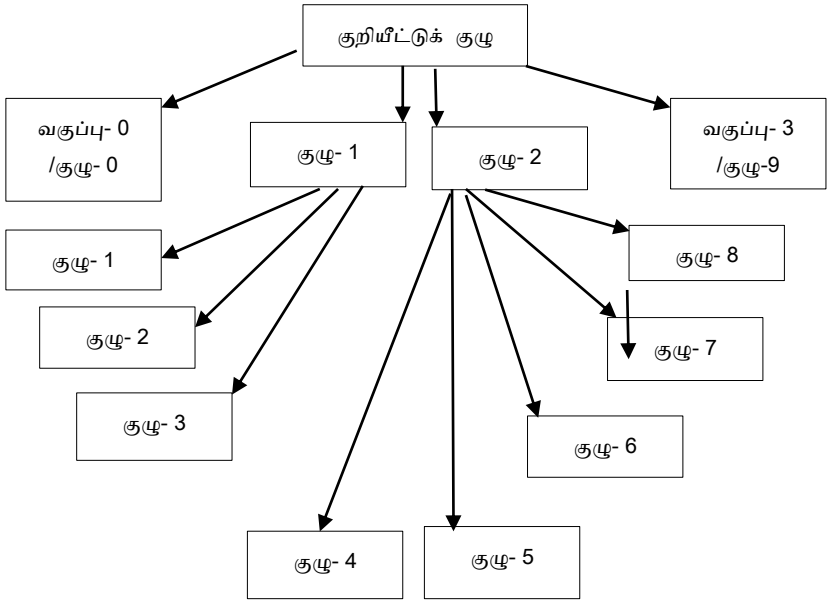
படம் 7 ஆ: 60 x 75 இயல்பாக்குதல் அளவுக்குப் பிறகு சின்னம்

படம் 7 இ: 60 x 75 மெல்லிய சின்னம்

சின்னம் அறிதல்

இந்திய மொழிகளில் ஒளிஎழுத்துணரியை (OCR) உருவாக்குவதில் மிகப்பெரிய சவால் எழுத்துக்குறி தொகுப்பின் பரந்த தன்மை மற்றும் அறியப்பட வேண்டிய சின்னங்களின் சிக்கலான வடிவியல் (complex geometry). சில தமிழ் எழுத்துக்கள் 2 அல்லது 3 துண்டிக்கப்பட்ட சின்னங்களால் ஆனவை, மேலும் இது தனிப்பட்ட அடையாளங்களை முதலில் அடையாளம் காணும் வகைப்பாடு மூலோபாயத்திற்கு அழைப்பு விடுகிறது, மேலும் அடுத்தடுத்த கட்டத்தில், அந்தக் எழுத்தைக் கண்டறிய பொருத்தமான அடுத்தடுத்த சின்னங்களின் எண்ணிக்கையை இணைக்கிறது. எழுத்துக்களில் 154 வெவ்வேறு சின்னங்கள் உள்ளன. இது அறிதல் நேரம் மற்றும் வகைப்படுத்தியின் சிக்கலை அதிகரிக்கிறது. எழுத்துக்களைச் சில கொத்துகளாகப் கூறிடுவது விரும்பத்தக்கது, இதனால் அறிதலின் போது தேடல் இடம் குறைவாக இருக்கும்; இதன் விளைவாக அறிதல் நேரம் குறைவாக இருக்கும். இது குழப்பம் அல்லது தவறான கிளஸ்டரிங்கின் நிகழ்தகவையும் குறைக்கிறது. தமிழ் எழுத்துக்களுக்கான ஒரு கிளைக் கட்டமைக்கப்பட்ட வகைப்பாடு திட்டம் அறிமுகப்படுத்தப்பட்டுள்ளது. இந்தி, பெங்காலி, கன்னடம் மற்றும் மராத்தி போன்ற பிற இந்திய மொழிகளுக்கும் இதேபோன்ற கட்டமைப்பை வடிவமைக்க முடியும்.

படம் 8: வகைப்படுத்தியின் கிளை அமைப்பு



தமிழில் மூன்று வெவ்வேறு நிலை வகைப்பாடுகள் உள்ளன

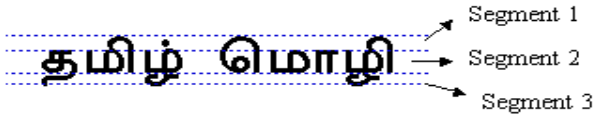
- உயரத்தின் அடிப்படையில் வகைப்பாடு
- மாத்திரைகள் / நீட்டிப்புகளின் அடிப்படையில் வகைப்பாடு
- மூன்றாம் மட்டத்தில் அறிதல்.

உயரத்தின் அடிப்படையில் வகைப்பாடு

தமிழில் உள்ள அனைத்து எழுத்துக்களும் உரை வரியின் 2 பகுதியை ஆக்கிரமித்துள்ளன. மேலும், வழக்கமான தமிழ் உரையில் தோன்றும் 60% எழுத்துக்கள் தரப்பட்டுள்ள பகுப்பாய்விலிருந்து வகுப்பு-0 க்கு சொந்தமானவை, அங்கு எழுத்துக்கள் கூறு-2க்கு மட்டுமே வரையறுக்கப்பட்டுள்ளன. எனவே, ஒரு உரை வரியின் கிடைமட்ட முந்திட்ட தோற்றவடிவம் எடுக்கப்பட்டால், இந்தக் கூறின் எந்த வரிசையின் திட்டமும் 1 அல்லது 3 கூறுகளுக்குச் சொந்தமான வரிசையுடன் ஒப்பிடும்போது மிக உயர்ந்த மதிப்பைக் கொண்டுள்ளது. தோற்றவடிவின் கூர்மையான உயர்வு மாற்றத்தைக் குறிக்கிறது கூறு-1 முதல் கூறு-2 வரை.

இதேபோல், தோற்றவடிவ மதிப்பில் கூர்மையான வீழ்ச்சி கூறு-2 இலிருந்து கூறு-3 க்கு மாறுவதோடு தொடர்புடையது. முந்திட்ட தோற்றவடிவின் மதிப்பில் மேலே குறிப்பிட்ட கூர்மையான மாற்றங்களைக் கண்டறிவதன் மூலம் வரி எல்லைகளைக் கண்டறிய முடியும். 3 வெவ்வேறு பிரிவுகளைக் கொண்ட எந்த தமிழ் உரையின் உரை வரியும் கீழே உள்ள படம் 9-இல் காட்டப்பட்டுள்ளது. ஒரு குறிப்பிட்ட சின்னத்தால் ஆக்கிரமிக்கப்பட்ட கூறுகள் நிலையானவை மற்றும் எழுத்துரு வேறுபாடுகளுக்குப் பொதுவாக மாறாதவை என்பதால், இந்த பிரிவுகளின் ஆக்கிரமிப்பைப் பொறுத்து ஒரு சின்னம் நான்கு வெவ்வேறு வகுப்புகளில் ஒன்றோடு தொடர்புடையது:

படம் 9: 4-வரி கூறிடலின் வெளியீடு



மாதிரைகள் / நீட்டிப்புகளின் அடிப்படையில் வகைப்பாடு

இந்த வகைப்படுத்தல் 1 மற்றும் 2 வகுப்புகளின் குறியீடுகளுக்கு மட்டுமே பொருந்தும், அவை மேல்நோக்கி (மாத்திராக்கள்) மற்றும் கீழ்நோக்கி நீட்டிப்புகளைக் கொண்டுள்ளன. இவை மேலும் குழுக்களாக வகைப்படுத்தப்படுகின்றன, அவை எழுத்தில் இருக்கும் ஏறுபவர்கள் மற்றும் இறங்குபவர்களைப் பொறுத்து இருக்கும். இந்த வகைப்பாடு அம்சம் அடிப்படையிலானது, அதாவது சோதனைச் சின்னத்தின் அம்ச திசையன்கள் இயல்பாக்கப்பட்ட பயிற்சி தொகுப்பின் அம்ச திசையன்களுடன் ஒப்பிடப்படுகின்றன. இந்த மட்டத்தில் பயன்படுத்தப்படும் அம்சங்கள் இரண்டாவது வரிசை வடிவியல் தருணங்கள் மற்றும் பயன்படுத்தப்பட்ட வகைப்படுத்தி அருகிலுள்ள அண்டை வகைப்படுத்தியாகும்.

படம் 10: வகுப்பு 1இன் துணைக்குழுக்கள்

படம் 10 அ: குழு 6 படம் 10 ஆ: குழு 7 படம் 10 இ: குழு 7

கெ கே கொ கி ணி பி னி யி னீ மீ பீ ணீ

படம் 11: வகுப்பு 1இன் துணைக்குழுக்கள்

படம் 11 அ: குழு 1 படம் 11 ஆ: குழு 5 படம் 11 இ: குழு 2

அ பு யு வு ஆ த ந ற ர

படம் 11 ஈ: குழு 3 படம் 11 இ: குழு 3 படம் 11 உ குழு 3

கு ரு சூ மு ழு பூ யூ று நு னு ணு

படம் 11 இ: குழு 3 படம் 11 இ: குழு 3

றூ தூ ஒ ழ

படம் 11 அ: குழு 1 படம் 11 பி: குழு 5 படம் 11 சி: குழு 2

படம் 11 டி: குழு படம் 11 இ: குழு 3

படம் 11: வகுப்பு 2 இன் துணைக்குழுக்கள்

மூன்றாம் மட்டத்தில் அறிதல்

மூன்றாவது மட்டத்தில், பண்புக்கூறு சார்ந்த அறிதல் செய்யப்படுகிறது. ஒவ்வொரு குழுவிற்கும், குறியீட்டு இயல்பாக்குதல் திட்டம் வேறுபட்டது. பண்புக்கூறு திசையனின் பரிமாணங்கள் வெவ்வேறு குழுக்களுக்கு வேறுபடுகின்றன, ஏனெனில் அவற்றின் இயல்பாக்க அளவுகள் வேறுபட்டவை. துண்டிக்கப்பட்ட டி.சி.டி குணகங்கள் (truncated DCT coefficients) அம்சங்களின்/பண்புக்கூறுகளின் இரண்டாவது தொகுப்பாகப் பயன்படுத்தப்படுகின்றன. சின்னங்களின் வகைப்பாட்டிற்கு அருகிலுள்ள அண்டை வகைப்படுத்தி பயன்படுத்தப்படுகிறது. பயிற்சி தொகுப்பில் சின்னம் என்ற அடையாளத்துடன் சின்னம் பெயரிடப்பட்டுள்ளது, இது அம்ச இடத்தில் அதற்கு மிக அருகில் உள்ளது. டி.சி.டி பண்புக்கூறுகளின் விஷயத்தில் யூக்ளிடியன் தூரம் பயன்படுத்தப்படுகிறது. ஒரு சின்னம் அதன் அருகிலுள்ள அண்டை ஒரு குறிப்பிட்ட எல்லைக்கு அப்பால் இருந்தால் தெரியவில்லை என்று அறிவிக்கப்படுகிறது.

பிந்தைய செயலாக்கம் (Postprocessing)

படம் 12 இல் கீழே காட்டப்பட்டுள்ள சில குழப்பமான எழுத்துக்கள் இன்னும் ஒரு நிலைக்குச் செல்கின்றன, அதில் குழப்பத்தைத் தீர்க்க சில ஹியூரிஸ்டிக்ஸ் பயன்படுத்தப்படுகின்றன.

படம் 12: குழப்பமான எழுத்துக்கள்

ஓ	ஒ
ல	வ
ற	ந
ஜ	ஜ
ம	ய
ன	ள

பயிற்சி தொகுப்பு (Training set)

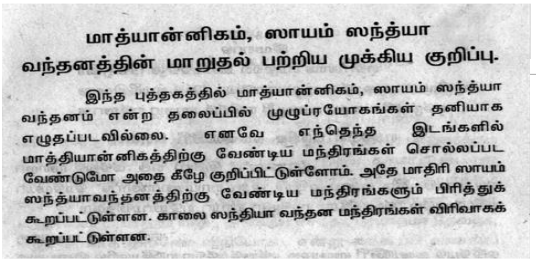
நல்ல அறிதல்/உணர்தல் துல்லியத்தைப் பெறுவதற்காக, 4000 மாதிரிகளைத் தாண்டிய அளவு கொண்ட பயிற்சித் தொகுப்பின் பரந்த தரவுத்தளத்தளம் உருவாக்கப்பட்டுள்ளது. ஒவ்வொரு எழுத்திலும் பல்வேறு இதழ்கள், நாவல்கள் மற்றும் தொழில்நுட்ப ஆவணங்கள் மற்றும் பல்வேறு தமிழ் நூல்களிலிருந்து சேகரிக்கப்பட்ட 25 முதல் 50 மாதிரிகள் உள்ளன. தரவுத்தளமானது தடிமனான மற்றும் சாய்வு எழுத்துக்களையும் உள்ளடக்கியது, அதே போல் கால்புள்ளி, அரைப்புள்ளி, கோலன் மற்றும் எண்கள் போன்ற சிறப்பு சின்னங்களையும் கொண்டுள்ளது. கம்பன், முரசு அஞ்சல் மற்றும் ஐ.எல்.ஏ.பி (ஐலீப்) போன்ற தமிழ் உரைப் பதிப்பான்களால் வழங்கப்பட்ட TM-TT வள்ளுவர், டிஏபி_அருள்மதி, இணைமதி, டிஎம்-டிடி பாரதி மற்றும் டிஏஎம்-அனிஷாய் போன்ற எழுத்துருக்கள் ஆகியவை சேர்க்கப்பட்டுள்ளன. கணினியைச் சோதிப்பதில் எழுத்துரு அளவுகளை 14 முதல் 20 வரை கையாளப்பட்டுள்ளது.

பயிற்சிப் பண்புக்கூறுத் தொகுப்பில் இயல்பாக்கப்பட்ட மற்றும் மெல்லிய சின்னங்களிலிருந்து பெறப்பட்ட பண்புக்கூறுகள் மற்றும் எழுத்தை அடையாளம் காண ஒரு லேபிள் ஆகியவை உள்ளன. அறியப்படாத சின்னத்தின் பண்புக்கூறுகள் பயிற்சி தொகுப்பில் அறியப்பட்ட சின்னங்களின் பண்புக்கூறுகளுடன் ஒப்பிடப்படுகின்றன மற்றும் பயிற்சி தொகுப்பில் நெருக்கமாக பொருந்தக்கூடிய ஒரு லேபிள் சோதனைத் தன்மைக்கு ஒதுக்கப்படுகிறது.

வகைப்பாடு முடிவுகள்

ஒரு நல்ல ஆவணத்தின் ஒட்டுமொத்த அறிதல் துல்லியம் (கணிசமாகக் குறைந்த அளவு சத்தத்துடன்) 98% ஆகும். சில முடிவுகள் கீழே 13 மற்றும் 14 புள்ளிவிவரங்களில் காட்டப்பட்டுள்ளன.

படம் 13 அ: அசல் ஆவணம் **படம் 13 பி: அறியப்பட்ட ஆவணம்**



மாத்யான்னிகம், ஸாயம் ஸந்த்யா
வந்தனத்தின் மாறுதல் பற்றிய முக்கிய குறிப்பு.

இந்த புத்தகத்தில் மாத்யான்னிகம், ஸாயம் ஸந்த்யா வந்தனம் என்ற தலைப்பில் முழுப்ரயோகங்கள் தனியாக எழுதப்படவில்லை. எனவே எந்தெந்த இடங்களில் மாத்தியான்னிகத்திற்கு வேண்டிய மந்திரங்கள் சொல்லப்பட வேண்டுமோ அதை கீழே குறிப்பிட்டுள்ளோம். அதே மாதிரி ஸாயம் ஸந்த்யாவந்தனத்திற்கு வேண்டிய மந்திரங்களும் பிரித்துக் கூறப்பட்டுள்ளன. காலை ஸந்தியா வந்தன மந்திரங்கள் விரிவாகக் கூறப்பட்டுள்ளன.

மாத்யான்னிகம், ஸாயம் ஸந்த்யா
வந்தனத்தின் மாறுதல் பற்றிய முக்கிய குறிப்பு.
இந்த புத்தகத்தில் மாத்யான்னிகம், ஸாயம் ஸந்த்யா
வந்தனம் என்ற தலைப்பில் முழுப்ரயே..ா.கங்கள் தனியாக
எழுதப்படவில்லை. எனவே எந்தெந்த இடங்களில்
மாத்தியான்னிகத்திற்கு வேண்டிய மந்திரங்கள் சொல்லப்பட
வேண்டுமோ அதை கீழே குறிப்பிட்டுளீளாம். அதே மாதிரி ஸாயம்
ஸ:ந்த்.யாவந்த:னத்திற்கு வேண்டிய மந்திரங்களும் பிரித்துக்
கூறப்பட்டு.ள்ளன. காலை ஸந்தியா வந்தன மந்திரங்கள் விரிவாகக் ,'
கூற்.பட்டு.,ளன. |

அரிதல் துல்லியம்

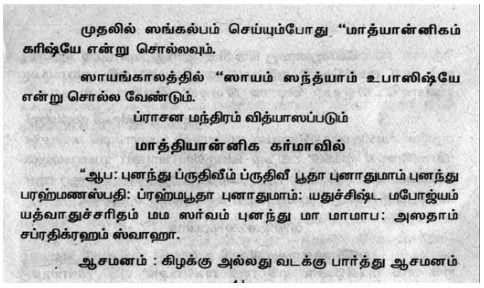
சின்னங்களின் மொத்த எண்ணிக்கை = 351

நிராகரிக்கப்பட்டது = 1 (~ அடையாளம்)

வகைப்படுத்தப்படாத = 2

அறிதல் விகிதம் = 99.48%

படம் 14 அ: அசல் ஆவணம் **படம் 14 பி: அங்கீகரிக்கப்பட்ட ஆவணம்**



முதலில் ஸங்கல்பம் செய்யும்போது "மாத்யான்னிகம் கரிஷ்யே என்று சொல்லவும்.

ஸாயங்காலத்தில் "ஸாயம் ஸந்த்யாம் உபாஸிஷ்யே என்று சொல்ல வேண்டும்.

ப்ராசன மந்திரம் வித்யாஸப்படும்

மாத்தியான்னிக கர்மாவில்

"ஆப: புனந்து ப்ருதிவீம் ப்ருதிவீ பூதா புனாதுமாம் புனந்து பரஹ்மணஸ்பதி: ப்ரஹ்மபூதா புனாதுமாம்: யதுச்சிஷ்ட மபோஜ்யம் யத்வாதுச்சரிதம் மம ஸர்வம் புனந்து மா மாமாப: அஸதாம் சப்ரதிக்ரஹம் ஸ்வாஹா.

ஆசமனம் : கிழக்கு அல்லது வடக்கு பார்த்து ஆசமனம்

முதவில் ஸங்கல்பம் செய்யும்போது "மாத்யான்னிகம்
கரிஷ்யே என்று சொல்லவும்.
ஸாயங்காலத்தில் "ஸாயம் ஸந்த்யாம் உபாஸிஷ்யே ,
என்று சொல்ல வேண்டும்.
,ப்ராசன மந்திரம் வித்யாஸப்படும்
மாத்தியான்னிக கர்மாவில்
"ஆப:. புனந்து ப்ருதிவீம் ப்ருதிவீ பூதா புனாதுமாம் புனந்து
பரஹ்மணஸ்பதி:. ப்ரஹ்மபூதா புனாதுமாம்:. யதுச்சிஷ்ட மபோஜ்யம்
யத்வாதுச்சரிதம் மம ஸர்வம் புனந்து மா மாமாப.: அஸதாம்
சப்ரதிக்ரஹம் ஸ்வாஹா.
, ஆசமனம் :. கிழக்கு அல்லது வடக்கு பார்த்து ஆசமனம் '
. .

உணர்தல் துல்லியம்

சின்னங்களின் மொத்த எண்ணிக்கை = 330

நிராகரிக்கப்பட்டது = 0

வகைப்படுத்தப்படாதது = 2

உணர்தல் விகிதம் = 99.39%

முடிவுரை

இங்கு ஒரு முழுமையான தமிழ் ஒளிவழி எழுத்துணரி (OCR) முறைமையை உருவாக்கும் முயற்சி பற்றி கூறப்பட்டுள்ளது, இது எழுத்துரு சுதந்திரனமான மற்றும் அளவு சுதந்திரமான சூழ்நிலையில் செயல்படுகிறது. ஒரு துல்லியமான வளைவு கண்டறிதல் அணுகுமுறை

பயன்படுத்தப்பட்டுள்ளது. இது உண்மையான மதிப்பைப் பற்றி ± 0.06 of வரம்பிற்குள் வளைவு கோணத்தைக் கண்டறிய உதவுகிறது. அளவீட்டு விளைவுகளைத் தவிர்ப்பதற்கும், பாத்திரத்தில் இருக்கும் சிதைவுகளைக் குறைப்பதற்கும் ஒரு சாம்பல் நிலை படத்தில் வளைவு சுழற்சி செய்யப்படுகிறது. வடிவியல் தருணங்கள் மற்றும் தனித்துவமான கொசைன் உருமாற்றங்கள் ஆகியவை பயன்படுத்தப்படும் பண்புக்கூறுகள் மற்றும் பயன்படுத்தப்படும் வகைப்படுத்தி அருகிலுள்ள அக்கம். ஒட்டுமொத்த அறிதல் துல்லியம் சுமார் 98% ஆகும்.

துணை நூல்கள்

"Optical character recognition." From Wikipedia, the free encyclopedia. Downloaded on 10.11.2020.

Aparna K.H. and V.S. Chakravarthy. 2003. "A complete OCR system development of Tamil magazine documents," In: Tamil Internet, Chennai, India, pp. 45-51, 2003.

Aparna K G and A G Ramakrishnan, "Tamil Gnaani - A complete Tamil OCR on windows", Proc. Tamil Internet 2001, Kuala Lumpur, Malaysia, Aug. 26-28, 2001, pp. 60-63.

Aparna K G and A G Ramakrishnan. "A Complete Tamil Optical Character Recognition System." Downloaded on 10.11.2020.

Chamila Liyanage. Thilini Nadungodage, Ruvan Weerasinghe. 2015. Developing a commercial grade Tamil OCR for recognizing font and size independent text. Conference Paper · August 2015. DOI: 10.1109/ICTER.2015.7377678.

Dhanya D and A G Ramakrishnan, "Simultaneous Recognition of Tamil and Roman Scripts", Proc. Tamil Internet 2001, Kuala Lumpur, Aug 26-28, 2001, pp. 64-68.

Dhanya, D.; A G Ramakrishnan and Peeta Basa Pati "Script Recognition in Bilingual Documents", Sadhana, Vol. 27 Part I, pp. 73-82, Feb. 2002.

Gonzalez R C and R E Woods, Digital Image Processing. Addison – Wesley, Massachusetts, 1993.

Gonzalez R C and R E Woods, Digital Image Processing. Addison – Wesley, Massachusetts, 1993.

Dhanya, D.; A G Ramakrishnan and Peeta Basa Pati “Script Recognition in Bilingual Documents”, Sadhana, Vol. 27 Part I, pp. 73-82, Feb. 2002.

O Trier, A K Jain and T Taxt, “Feature extraction methods for character recognition – a survey” Vol. 29. Pattern Recognition, pp. 641-662, 1996.

Ramakrishnan, A.G. 2000. A complete OCR for printed Tamil text. Conference Paper · July 2000, DOI: 10.13140/RG.2.1.4593.5209

Ramanathan, R.; S. Ponmathavan, N. Valliappan, L. Thaneshwaran, A.S. Nair and K.P. Soman, 2009. “Optical Character Recognition for English and Tamil Using Support Vector Machines,” In: Advances in Computing, Control, & Telecommunication Technologies, ACT '09, International Conference on, pp. 610-612, IEEE 2009.

Suresh, R.M., Arumugan, S., and Aravanan, K.P. 2000. “Recognition of Handwritten Tamil characters using fuzzy classificatory approach”, in Proceedings of the Tamil Internet 2000 Conference, Singapore.

வினா வங்கி

1.பொருத்தமான விடைத் தேர்ந்தெடுத்தல் (15 வினாக்கள்)

1) எழுத்துணரியின் தொடக்கம்

அ.1860க்குப் பிந்தையது.

ஆ. 1870க்கு முந்தையது.

இ. 1870க்கு பிந்தையது.√

ஈ. 1860இல்.

2) ஒளி எழுத்துணரியில் வர்ணம் முக்கியமானதாகும். ஏனென்றால்

அ. டிராப் வர்ணத்தை உபயோகிப்பது வருடியின் வெளியீட்டின் அளவை குறைக்கும்.

√ஆ. டிராப் வர்ணத்தை உபயோகிப்பது வருடியின் வெளியீட்டின் அளவைக்குறைக்கும் துல்லியத்தை அதிகரிக்கும்.

இ. டிராப் வர்ணத்தை உபயோகிப்பது வருடியின் வெளியீட்டின் துல்லியத்தை அதிகரிக்கும்.

ஈ. டிராப் வர்ணத்தை உபயோகிப்பது வருடியின் வெளியீட்டின் அளவைக் அதிகரிக்கும் துல்லியத்தை குறைக்கும்.

3) OCR பொதுவாக

அ. ஒரு "ஆன்ப்லைன்" செயல்முறையாகும், இது ஒரு நிலையான ஆவணத்தை பகுப்பாய்வு செய்கிறது.

√ஆ. ஒரு "ஆஃப்லைன்" செயல்முறையாகும், இது ஒரு நிலையான ஆவணத்தை பகுப்பாய்வு செய்கிறது.

இ. ஒரு "ஆஃப்லைன்" செயல்முறையாகும், இது ஒரு இயக்க ஆவணத்தை பகுப்பாய்வு செய்கிறது.

ஈ. ஒரு "ஆன்லைன்" செயல்முறையாகும், இது ஒரு இயக்க ஆவணத்தை பகுப்பாய்வு செய்கிறது.

4) இருமையாக்கம் (Binarisation)

√அ. ஒரு படத்தை வண்ணம் அல்லது கிரேஸ்கேலில் இருந்து கருப்பு மற்றும் வெள்ளைக்கு மாற்றும்

ஆ. நேர்மறை மற்றும் எதிர்மறை புள்ளிகள், மென்மையான விளிம்புகளை அகற்றும்.

இ. கிளிஃப் அல்லாத பெட்டிகளையும் வரிகளையும் சுத்தம் செய்கிறது

ஈ. நெடுவரிசைகள், பத்திகள், தலைப்புகள் போன்றவற்றை அடையாளம் காட்டுகிறது.

2. பொருத்துக

1) அ. தட்டச்சு செய்யப்பட்ட உரை, ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது - நுண்ணறிவு சொல் அங்கீகாரம் (IWR)

ஆ. தட்டச்சு செய்யப்பட்ட உரையை குறிவைக்கிறது – நுண்ணறிவு எழுத்து அறிதல் (ஐ.சி.ஆர்)

இ. கையால் எழுதப்பட்ட அச்சுப்பொறி அல்லது கர்சீவ் உரையை ஒரு நேரத்தில் ஒரு கிளிஃப் அல்லது எழுத்தை குறிவைக்கிறது, பொதுவாக இயந்திர கற்றல் இதில் அடங்கும் – ஆப்டிகல் சொல் அறிதல்

ஈ. கையால் எழுதப்பட்ட அச்சுப்பொறி அல்லது கர்சீவ் உரையையும் குறிவைக்கிறது - ஒளி எழுத்துணரி (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் (ஓ.சி.ஆர்))

2) முன் செயலாக்க நுட்பங்கள் - பொருத்துக

அ. ஸ்கேன் செய்யும் போது ஆவணம் சரியாக சீரமைக்கப்படாவிட்டால், உரையின் வரிகளை கிடைமட்டமாக அல்லது செங்குத்தாக மாற்ற சில டிகிரி கடிகார திசையில் அல்லது எதிரெதிர் திசையில் சாய்ந்து கொள்ள வேண்டியிருக்கும் – வரி நீக்கம் (Line removal)

ஆ. நேர்மறை மற்றும் எதிர்மறை புள்ளிகள், மென்மையான விளிம்புகளை அகற்றும் – இருமையாக்கம் (பைனரைசேஷன்/Binarisation)

இ. ஒரு படத்தை வண்ணம் அல்லது கிரேஸ்கேலில் இருந்து கருப்பு மற்றும் வெள்ளை நிறமாக மாற்றும் - டெஸ்பெக்கிள் (Despeckle)

ஈ. கிளிஃப் அல்லாத பெட்டிகளையும் வரிகளையும் சுத்தம் செய்கிறது – டி-ஸ்கேவ் (De-skew)

3) முன் செயலாக்க நுட்பங்கள் – பொருத்துக

அ. தளவமைப்பு பகுப்பாய்வு அல்லது "மண்டலம்" – ஒவ்வொரு எழுத்துக்குறி OCR-க்கு, படக் கலைப்பொருட்கள் காரணமாக இணைக்கப்பட்டுள்ள பல எழுத்துக்கள் பிரிக்கப்படுகிறது.

ஆ. வரி மற்றும் சொல் கண்டறிதல் - ஸ்கிரிப்டை அடையாளம் காண்கிறது.

இ. ஸ்கிரிப்ட் அறிதல் – சொல் மற்றும் எழுத்து வடிவங்களுக்கான அடிப்படைகளை நிறுவுகிறது, தேவைப்பட்டால் சொற்களைப் பிரிக்கிறது.

ஈ. எழுத்து தனிமைப்படுத்தல் அல்லது "கூறிடல்" – நெடுவரிசைகள், பத்திகள், தலைப்புகள் போன்றவற்றை தனித்துவமான தொகுதிகளாக அடையாளம் காட்டுகிறது.

4) பேச்சு அறிதல் வரலாறு – பொருத்துக

அ. மூன்று பெல் லேப்ஸ் ஆராய்ச்சியாளர்கள், ஸ்டீபன் பாலாஷேக், ஆர். பிதுல்ப், மற்றும் கே. எச். டேவிஸ் ஆகியோர் ஒற்றை பேச்சாளர் இலக்க அறிதலுக்காக "ஆட்ரி" என்ற அமைப்பை உருவாக்கினர். -1962

ஆ. குன்னர் ஃபான்ட் பேச்சு உற்பத்தியின் மூல-வடிகட்டி மாதிரியை உருவாக்கி வெளியிட்டார். - 1952

இ. ஐபிஎம் தனது 16-வார்த்தை "ஷூ பாக்ஸ்" இயந்திரத்தின் பேச்சு அறிதல் திறனை உலக கண்காட்சியில் நிரூபித்தது - 1960

ஈ. 1966 - பேச்சு குறியீட்டு முறையான லீனியர் ப்ரிடிக்டிவ் கோடிங் (எல்பிசி) முதன்முதலில் நாகோயா பல்கலைக்கழகத்தின் புமிதாடா இட்டகுரா மற்றும் நிப்பான் டெலிகிராப் மற்றும் டெலிபோனின் (என்.டி.டி) ஷூசோ சைட்டோ ஆகியோரால் முன்மொழியப்பட்டது.

5) பேச்சு அறிதல் வரலாறு – பொருத்துக

அ. 1971 – லியோனார்ட் பாம் மார்கோவ் சங்கிலிகளின் கணிதத்தை உருவாக்கினார்

ஆ. 1972 – ICASSP பிலடெல்பியாவில் நடைபெற்றது,

இ. 1976 முதல் - மாசசூசெட்ஸின் நியூட்டனில் IEEE ஒலியியல், பேச்சு மற்றும் சிக்னல் செயலாக்கக் குழு ஒரு மாநாட்டை நடத்தியது.

ஈ. 1960களின் பிற்பகுதி – பேச்சு புரிந்துணர்வு ஆராய்ச்சிக்கு தர்பா ஐந்து ஆண்டுகள் நிதியளித்தது

6) சின்தசைசர் தொழில்நுட்பங்கள் - பொருத்துக

அ, பதிவுசெய்யப்பட்ட பேச்சின் பிரிவுகளின் ஒருங்கிணைப்பு (அல்லது ஒன்றாக சரம்) அடிப்படையாகக் கொண்டது. - டொமைன்-குறிப்பிட்ட தொகுப்பு

ஆ. பதிவு செய்யப்பட்ட பேச்சின் பெரிய தரவுத்தளங்களைப் பயன்படுத்துகிறது. - டிஃபோன் தொகுப்பு

இ. ஒரு மொழியில் நிகழும் அனைத்து டிஃபோன்களையும் (ஒலி-க்கு-ஒலி மாற்றங்கள்) கொண்ட குறைந்தபட்ச பேச்சு தரவுத்தளத்தைப் பயன்படுத்துகிறது.- அலகுத் தேர்வு தொகுப்பு (Unit selection synthesis)

ஈ. முழுமையான சொற்களை உருவாக்க முன்பே பதிவுசெய்யப்பட்ட சொற்களையும் சொற்றொடர்களையும் இணைக்கிறது. ஒத்திசைவு தொகுப்பு (Concatenative synthesis) -

3. சரியா/தவறா (15 வினாக்கள்)

1) ஒளி எழுத்துணரி உருவக்க நுட்பமான உணர்திறன் வாய்ந்த வருடி தேவைப்படும்.

√சரி/தவறு

2) படச்செயலாக்கம் என்பது சில வழிமுறைகளால் படங்களின் தரத்தை குறைக்கும் செயல்முறையாகும்.

சரி/தவறு√

4. ஒரு சொல்/சொற்றொடரில் விடைதருக (15 வினாக்கள்)

1) ஒளி எழுத்துணரி செயல்பாட்டிற்கான ஒரு அனுமானத்தைக் கூறுக.

2) கரடுமுரடான வளைவு மதிப்பீட்டில் சம்பந்தப்பட்ட ஒரு படியைக் கூறுக.

3) நன்றாக வளைவு கண்டறிதல் உள்ளடக்கும் ஒரு படியைக் கூறுக.

4) எழுத்துக் கூறிடல் என்றால் என்ன?

5) சின்னம் முன் செயலாக்கம் என்றால் என்ன?

5. ஒரு பத்தியில் விடை தருக (10 வினாக்கள்)

1) படச்செயலாக்கம் பற்றி விளக்குக.

2) தமிழ் எழுத்துக்களின்பண்புகளை விளக்குக.

3) ஒளி எழுத்துணரி உருவாக்கத்தில் முன் செயலாக்கம் என்றால் என்ன?

4) உயர்த்தின் அடிப்படையில் வகைப்படுத்தல் என்றால் என்ன?

5) சின்னம் முன் செயலாக்கம் பற்றி கூறுக.

6. மூன்று பக்க அளவில் விடை தருக?

1) தமிழுக்கான ஒளி எழுத்துணரி உருவாக்கம் பற்றி கட்டுரை வரைக.

3. தமிழுக்கு கூகுள் எழுத்துணரி தொழில் நுட்பம்

ஒளிவழி எழுத்துணரி (Optical Character Recognition (OCR) என்பது ரியாலிட்டி உலகத்தையும் மெய்நிகர் சொல்லையும் இணைப்பதற்கான வழிகளில் ஒன்றாகும். முதல் ஒளிவழி எழுத்துணரி அமைப்பு 1920களின் பிற்பகுதியில் அறிமுகப்படுத்தப்பட்டது. ஒளிவழி எழுத்துணரியின் நோக்கம் படத்திலிருந்து உரையை அறிந்துகொள்வதாகும். இருப்பினும், நிறைய காரணிகளால் மிக உயர்ந்த துல்லியத்தை அடைவது மிகவும் சவாலானது. இந்தச் சிக்கலைச் சமாளிக்க கூகிள் கிளவுட் விஷன் ஏபிஐ ஒன்றில் கூகிள் எவ்வாறு தீர்வை உருவாக்குகிறது என்பது இங்கு அறிமுகப்படுத்தப்படுகிறது.

கூகிள் ஒளிவழி எழுத்துணரி என்பது கூகிள் வழங்கிய ஒரு பயன்பாடாகும், இதன் பயனர்கள் அதன் மேடையில் ஒளிவழி எழுத்துணரி தொழில்நுட்பத்தை அணுகலாம் மற்றும் அவற்றின் ஸ்கேன் செய்யப்பட்ட படங்கள் அல்லது PDF கோப்புகளை உரை திருத்தக்கூடிய கூகிள் செயல்பாடாக மாற்றலாம். விரும்பிய கோப்பு மாற்றப்பட்டதும், இறுதி தயாரிப்பு விரும்பியபடி பயன்படுத்தப்படலாம். இதைக் கூகுள் ஆவணத்தில் திருத்தலாம், மற்றவர்களுடன் பகிர்வதற்கான மின்னஞ்சல்களில் இணைக்கப்படலாம் அல்லது கூடுதல் பயன்பாட்டிற்கு பதிவிறக்கம் செய்யலாம்.

கூகிள் ஒளிவழி எழுத்துணரியை எவ்வாறு பயன்படுத்துவது

கூகிள் ஒளிவழி எழுத்துணரியைப் பயன்படுத்த ஒரு தனிநபருக்குத் தேவையானது ஸ்கேனர் மற்றும் கூகிள் கணக்கு ஐடி மட்டுமே. நபர் முதலில் தேவையான ஆவணத்தை ஸ்கேன் செய்து கணினியில் சேமிக்கிறார். அந்த நபர் தங்கள் கணக்கு ஐடியைப் (account ID) பயன்படுத்தி கூகிள் செயல்பாட்டில் உள்நுழைகிறார். அவர்களின் கணக்குப் பக்கம் பின்னர் காட்டப்படும். பதிவேற்ற கோப்பு பொத்தானைக் கிளிக் செய்தால் மாற்று அமைப்புகளை சரிசெய்ய ஒரு சாளரம் தோன்றும்.

பயனர் 'PDF மற்றும் பிற படக் கோப்புகளிலிருந்து உரையை கூகுள் ஆவணங்களுக்கு மாற்று' விருப்பத்திற்கு எதிராக கொடுக்கப்பட்ட பெட்டியை சரிபார்க்க வேண்டும். இது முடிந்ததும், அவர்கள் பதிவேற்ற பொத்தானைக் கிளிக் செய்து மாற்ற வேண்டிய அந்தந்த கோப்பைத் தேர்வு செய்கிறார்கள். கூகிள் டாக் (Google Doc) பின்னர் கோப்பை மாற்றி, மாற்று படத்தை அசல் படத்திற்கு கீழே காண்பிக்கும். ஆவணம் இப்போது எடிட்டிங் நோக்கங்களுக்காக கிடைக்கிறது. வேறு எந்த மின்னணு ஆவணத்தையும் நாம் பயன்படுத்துவதைப் போலவே இந்த ஆவணத்தையும் பயன்படுத்தலாம். ஒருவர் ஆவணத்தின் நகலை உருவாக்கலாம், அதைத் திருத்தக்கூடிய வடிவத்தில் பதிவிறக்கம் செய்யலாம், மற்றவர்களுடன் மின்னஞ்சலில் பகிரலாம் அல்லது வலையில் வெளியிடலாம். இறுதிக் கோப்பு கூகுள் டாக்ஸ் கோப்பு கோப்புறைகளிலும் சேமிக்கப்படும், மேலும் தேவைப்படும்போது அணுகலாம்.

கூகிள் ஒளிவழி எழுத்துணரியிலிருந்து பயன்கிடைகிறது

கூகிள் ஒளிவழி எழுத்துணரி பதிவிறக்கம் செய்ய பலர் விரும்பவில்லை, ஏனெனில் மென்பொருள் தங்கள் கணினி வன்வட்டில் அதிக இடத்தை எடுக்கும் என்று அவர்கள் அஞ்சுகிறார்கள். மற்றவர்கள் வணிக நோக்கங்களுக்காகத் தொடர்ந்து பயணம் செய்கிறார்கள், மேலும் ஸ்கேன் செய்யப்பட்ட ஆவணங்களை மாற்ற ஒளிவழி எழுத்துணரி தேவைப்படுவதைக் காணலாம். கூகிள் ஒளிவழி எழுத்துணரி என்பது அவர்களின் தனிப்பட்ட பயன்பாட்டிற்காக ஒளிவழி எழுத்துணரி மென்பொருளை வாங்குவதில் பணம் செலவழிக்க விரும்பாத அனைவருக்கும் ஒரு வரம். Web.desired-ஐ அணுகுவதன் மூலம் அவர்கள் இப்போது பயன்பாட்டைப் பயன்படுத்தலாம்.

ஒளிவழி எழுத்துணரி (OCR) டுடோரியல்

கூகிள் மேகக்கணி இயங்குதளத்தில் ஒளிவழி எழுத்துணரி (OCR) எவ்வாறு செய்வது என்பதை அறிக. கூகிள் மேகக்கணி சேமிப்பகத்தில் படக் கோப்புகளை எவ்வாறு பதிவேற்றுவது, கூகிள் மேகக்கணி பார்வை API-ஐப் பயன்படுத்தி படங்களிலிருந்து உரையை பிரித்தெடுப்பது, கூகிள் மேகக்கணி மொழிபெயர்ப்பு API-ஐப் பயன்படுத்தி உரையை மொழிபெயர்ப்பது மற்றும் உங்கள் மொழிபெயர்ப்புகளை மேகக்கணி சேமிப்பகத்தில் எவ்வாறு சேமிப்பது என்பதை இந்த பயிற்சி நிரூபிக்கிறது. கூகிள் கிளவுட் பப்/சப் (Google Cloud Pub/Sub) பல்வேறு பணிகளை வரிசைப்படுத்தவும், சரியான கிளவுட் செயல்பாடுகளைச் செயல்படுத்தவும் பயன்படுகிறது.

குறிக்கோள்கள்

- பல பின்னணி மேகக்கணி செயல்பாடுகளை எழுதி வரிசைப்படுத்தவும்.
- படங்களை மேகக்கணி சேமிப்பகத்தில் பதிவேற்றவும்.
- பதிவேற்றிய படங்களில் உள்ள உரையை பிரித்தெடுக்கவும், மொழிபெயர்க்கவும் சேமிக்கவும்.

செலவுகள்

இந்த டுடோரியல் கிளவுட் பிளாட்ஃபார்மின் பில் செய்யக்கூடிய கூறுகளைப் பயன்படுத்துகிறது, அவற்றுள்:

- மேகக்கணி செயல்பாடுகள்
- கூகிள் கிளவுட் பப் /சப் (Pub/Sub)
- கூகிள் மேகக்கணி சேமிப்பிடம்
- கூகிள் மேகக்கணி மொழிபெயர்ப்பு API
- கூகிள் மேகக்கணி பார்வை API

உங்கள் திட்டமிடப்பட்ட பயன்பாட்டின் அடிப்படையில் செலவு மதிப்பீட்டை உருவாக்க விலை கால்குலேட்டரைப் பயன்படுத்தவும்.

புதிய கிளவுட் பிளாட்ஃபார்ம் பயனர்கள் இலவச சோதனைக்கு தகுதியுடையவர்களாக இருக்கலாம்.

நீங்கள் தொடங்கும் முன்

1. உங்கள் கூகிள் கணக்கில் உள்நுழைக.

   உங்களிடம் ஏற்கனவே ஒன்று இல்லையென்றால், புதிய கணக்கிற்கு பதிவுபெறுக.

2. கூகிள் மேகக்கணி கன்சோலில், திட்டத் தேர்வாளர் பக்கத்தில், கூகிள் மேகக்கணித் திட்டத்தைத் தேர்ந்தெடுக்கவும் அல்லது உருவாக்கவும்.

குறிப்பு: இந்த நடைமுறையில் நீங்கள் உருவாக்கும் வளங்களை வைத்திருக்க நீங்கள் திட்டமிடவில்லை என்றால், ஏற்கனவே இருக்கும் திட்டத்தைத் தேர்ந்தெடுப்பதற்குப் பதிலாக ஒரு திட்டத்தை உருவாக்கவும். இந்தப் படிகளை நீங்கள் முடித்த பிறகு, திட்டத்துடன் தொடர்புடைய அனைத்து வளங்களையும் நீக்கிவிட்டுத் திட்டத்தை நீக்கலாம்.

திட்ட தேர்வுக்குழு பக்கத்திற்குச் செல்லவும்

3. உங்கள் கிளவுட் திட்டத்திற்கு பில்லிங் இயக்கப்பட்டிருப்பதை உறுதிசெய்க. உங்கள் திட்டத்திற்கு பில்லிங் இயக்கப்பட்டிருப்பதை எவ்வாறு உறுதிப்படுத்துவது என்பதை அறிக.

4. கிளவுட் செயல்பாடுகள், கிளவுட் பில்ட், கிளவுட் பப் / சப், கிளவுட் ஸ்டோரேஜ், கிளவுட் டிரான்ஸ்லேஷன் மற்றும் கிளவுட் விஷன் ஏபிஐகளை இயக்கவும்.

API களை இயக்கவும்

5. கிளவுட் SDK ஐ நிறுவி துவக்கவும்.

நீங்கள் ஏற்கனவே கிளவுட் எஸ்.டி.கே நிறுவப்பட்டிருந்தால், பின்வரும் கட்டளையை இயக்குவதன் மூலம் அதைப் புதுப்பிக்கவும்:

gcloud கூறுகள் புதுப்பிப்பு

6. உங்கள் வளர்ச்சி சூழலைத் தயாரிக்கவும்.

Node.js பைதான் கோ ஜாவா

Node.js அமைவு வழிகாட்டிக்குச் செல்லவும்

தரவின் ஓட்டத்தைக் காட்சிப்படுத்துகிறது

OCR டுடோரியல் பயன்பாட்டில் தரவின் ஓட்டம் பல படிகளை உள்ளடக்கியது:

1. எந்த மொழியிலும் உரையைக் கொண்டிருக்கும் படம் கிளவுட் ஸ்டோரேஜில் (Cloud Storage) பதிவேற்றப்படுகிறது.

2. ஒரு கிளவுட் செயல்பாடு தூண்டப்படுகிறது, இது உரையை பிரித்தெடுக்க மற்றும் மூல மொழியைக் கண்டறிய விஷன் APIஐப் (Vision API) பயன்படுத்துகிறது.

3. ஒரு பப்/சப் (Pub/Sub) தலைப்புக்கு ஒரு செய்தியை வெளியிடுவதன் மூலம் உரை மொழிபெயர்ப்பிற்காக வரிசைப்படுத்தப்படுகிறது. மூல மொழியிலிருந்து வேறுபட்ட ஒவ்வொரு இலக்கு மொழிக்கும் ஒரு மொழிபெயர்ப்பு வரிசைப்படுத்தப்படுகிறது.

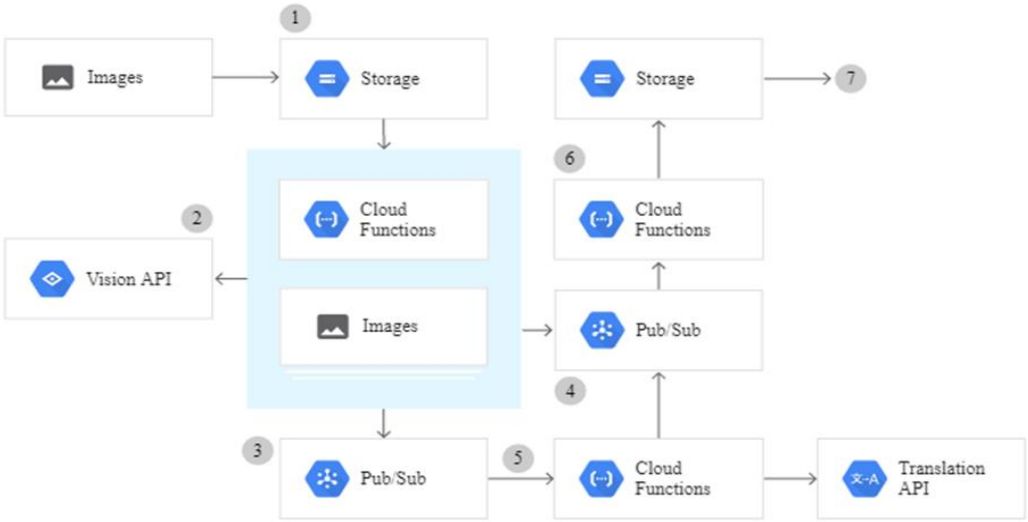
4. ஒரு இலக்கு மொழி மூல மொழியுடன் பொருந்தினால், மொழிபெயர்ப்பு வரிசை தவிர்க்கப்பட்டு, உரை முடிவு வரிசையில், மற்றொரு பப்/சப் (Pub/Sub) தலைப்புக்கு அனுப்பப்படும்.

5. ஒரு கிளவுட் செயல்பாடு மொழிபெயர்ப்பு வரிசையில் உள்ள உரையை மொழிபெயர்க்க மொழிபெயர்ப்பு API ஐப் பயன்படுத்துகிறது. மொழிபெயர்க்கப்பட்ட முடிவு முடிவு வரிசையில் அனுப்பப்படுகிறது.

6. மற்றொரு கிளவுட் செயல்பாடு, மொழிபெயர்க்கப்பட்ட உரையை முடிவு வரிசையில் இருந்து கிளவுட் ஸ்டோரேஜுக்கு சேமிக்கிறது.

7. முடிவுகள் ஒவ்வொரு மொழிபெயர்ப்பிற்கும் txt கோப்புகளாக கிளவுட் ஸ்டோரேஜில் காணப்படுகின்றன.

படிகளைக் காட்சிப்படுத்த இது உதவக்கூடும்:



விண்ணப்பத்தைத் தயாரித்தல்

1. படங்களை பதிவேற்ற ஒரு கிளவுட் ஸ்டோரேஜ் வாளியை உருவாக்கவும், அங்கு YOUR_IMAGE_BUCKET_NAME என்பது உலகளவில் தனித்துவமான வாளி பெயர்:

gsutil mb gs: // YOUR_IMAGE_BUCKET_NAME

உரை மொழிபெயர்ப்புகளைச் சேமிக்க கிளவுட் ஸ்டோரேஜ் வாளியை உருவாக்கவும், அங்கு YOUR_RESULT_BUCKET_NAME என்பது உலகளவில் தனித்துவமான வாளி பெயர்:

gsutil mb gs: // YOUR_RESULT_BUCKET_NAME

3. மொழிபெயர்ப்பு கோரிக்கைகளை வெளியிட கிளவுட் பப்/சப் (Pub/Sub) தலைப்பை உருவாக்கவும், அங்கு உங்கள் மொழிபெயர்ப்பு கோரிக்கை தலைப்பின் பெயர் YOUR_TRANSLATE_TOPIC_NAME:

gcloud pubsub தலைப்புகள் YOUR_TRANSLATE_TOPIC_NAME-ஐ உருவாக்குகின்றன

முடிக்கப்பட்ட மொழிபெயர்ப்பு முடிவுகளை வெளியிட கிளவுட் பப்/சப் (Pub/Sub) தலைப்பை உருவாக்கவும், அங்கு உங்கள் மொழிபெயர்ப்பு முடிவு தலைப்பின் பெயர் YOUR_RESULT_TOPIC_NAME:

gcloud pubsub தலைப்புகள் YOUR_RESULT_TOPIC_NAME-ஐ உருவாக்குகின்றன

உங்கள் உள்ளூர் கணினியில் மாதிரி பயன்பாட்டு களஞ்சியத்தை குளோன் செய்யுங்கள்:

Node.js பைதான் கோ ஜாவா

git clone https://github.com/GoogleCloudPlatform/nodejs-docs-samples.git

மாற்றாக, நீங்கள் மாதிரியை ஒரு ஜிப் கோப்பாக பதிவிறக்கம் செய்து பிரித்தெடுக்கலாம்.

மேகக்கணி செயல்பாடுகள் மாதிரிக் குறியீட்டைக் கொண்ட கோப்பகத்திற்கு மாற்றவும்:

Node.js பைதான் கோ ஜாவா

cd nodejs-docs-மாதிரிகள் / செயல்பாடுகள் / ocr / app /

கூகிள் வலை அடிப்படையிலான OCR சேவையின் ரகசியம்

டெசராக்ட் ஒளிவழி எழுத்துணரி

ஒளிவழி எழுத்துணரியைப் பற்றி பேசுகையில், டெசராக்ட் என்பது பிரபலமான திறந்த மூல நூலகங்களில் ஒன்றாகும், இது ஒளிவழி எழுத்துணரியை இயக்க அனைவருக்கும் பயன்படும். டெசராக்ட் ஹெச்பி மூலம் கண்டறியப்பட்டது மற்றும் மேம்பாடு 2006 ஆம் ஆண்டு முதல் கூகிள் நிதியுதவி அளித்துள்ளது. டெசராக்ட் 3. எக்ஸ் மாடல் பழைய பதிப்பாகும், 4. எக்ஸ் பதிப்பு ஆழ்ந்த கற்றல் (LSTM/எல்எஸ்டிஎம்) மூலம் கட்டப்பட்டுள்ளது. 3.x மற்றும் 4.x க்கு இடையிலான வித்தியாசத்தை நீங்கள் புரிந்து கொள்ள விரும்பினால், மேலும் விவரங்களுக்கு பகிர்வை நீங்கள் பார்வையிடலாம்.

C ++ ஆல் டெசராக்ட் செயல்படுத்தப்படுவதால், அதை மற்ற பைதான் நூலகமாக நாம் பயன்படுத்த முடியாது. உண்மையில், நாம் பைத்தானில் சி-ஏபிஐக்கு (C-API) அழைக்கலாம், ஆனால் அது பயனர் நட்பு அல்ல. எனவே, நம் வாழ்க்கையை எளிதாக்குவதற்காக பைதான் ரேப்பர் (python wrapper), பைடெசெராக்ட் (pytesserac) அறிமுகப்படுத்தப்பட்டுள்ளது.,

வரையறை

கட்டமைப்பு வடிவமைப்பை அறிமுகப்படுத்துவதற்கு முன், சில வரையறைகள் அறிமுகப்படுத்தப்பட வேண்டும்.

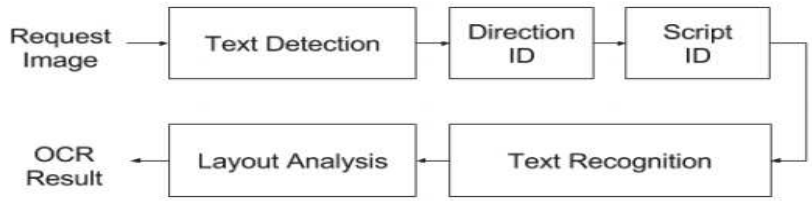
- ஸ்கிரிப்ட்க்கு எதிராக மொழி (Script vs Language): ஸ்கிரிப்ட் மொழியிலிருந்து வேறுபட்டது. ஸ்கிரிப்ட் எழுத்து முறையை குறிக்கிறது, அதே நேரத்தில் மொழி பேசும் மொழியை குறிக்கிறது. பின்வரும் படத்தில், “தரவு விஞ்ஞானி” என்பது ரோமானிய எழுத்தில் ஆங்கில மொழியாகும், “ஷுஜு கெக்சுஜியா” என்பது ரோமானிய எழுத்துக்களில் சீன மொழியாகும்.

ஸ்கிரிப்ட் Vs மொழி

- எல்லை பெட்டி (Bounding Box): பிற OCR அமைப்புகளிலிருந்து வேறுபட்டது, எல்லை பெட்டியில் ஒற்றை எழுத்துக்குறி அல்லது ஒற்றை வார்த்தைக்கு பதிலாக கண்டறியப்பட்ட உரையின் ஒற்றை வரி அடங்கும்.
- மாதிரி கருத்தில் (Model Consideration): துல்லியத்தைத் தவிர, செலவு, பொதுமயமாக்கல் மற்றும் பராமரித்தல் ஆகியவை மாதிரியை உருவாக்கக் கருதப்படுகின்றன.

Google மேகக்கணி பார்வை API

ஒரு படத்தைக் கொடுத்து, கூகிள் விஷன் APIஇல் இறுதி முடிவைப் பெற இது 5 நிலைகளைக் கடந்து செல்கிறது.



கூகிள் கிளவுட் விஷன் ஏபிஐ கட்டிடக்கலை (வாக்கர் மற்றும் பலர், 2018)

உரை கண்டறிதல் (Text Detection)

முதல் கட்டம் உரையின் வரிகளைக் கண்டறியவும் இடவெல்லைக்குட்படுத்தவும் மற்றும் ஒரு குழும எல்லைப் பெட்டிகளை உருவாக்கவும் .வழக்கமான நரம்பியல் நெட்வொர்க் (Conventional Neural Network (CNN/சி.என்.என்)) அடிப்படையிலான மாதிரியைப் பயன்படுத்துவதாகும்

திசை அடையாளம் (Direction Identification)

எல்லைக்குட்பட்ட பெட்டியின் திசையை வகைப்படுத்துகிறது. தேவைப்பட்டால், உரையாகத் தவறாகக் கண்டறியப்பட்டதால் சில எல்லை பெட்டி வடிகட்டப்படும்.

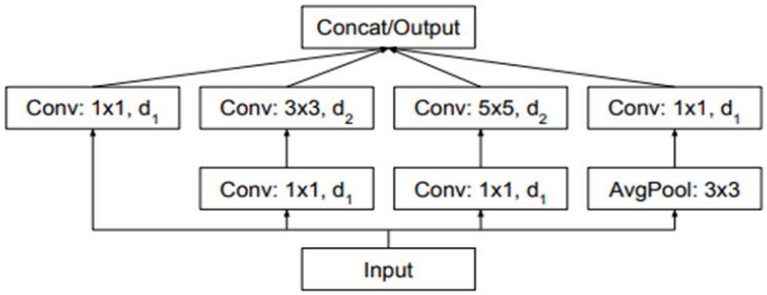
ஸ்கிரிப்ட் அடையாளம் (Script Identification)

எல்லைக்குட்பட்ட பெட்டிக்கு ஸ்கிரிப்டை அடையாளம் காட்டுகிறது. எல்லைக்குட்பட்ட பெட்டியில் ஒரு 1 ஸ்கிரிப்ட் இருப்பதாக கருதப்படுகிறது, ஆனால் ஒரு படத்திற்கு பல ஸ்கிரிப்ட்களை அனுமதிக்கிறது.

உரை அறிதல் (Text Recognition)

இது ஒளிவழி எழுத்துணரியின் முக்கிய பகுதியாகும், இது படத்திலிருந்து உரையை அறிகிறது. இது எழுத்து அடிப்படையிலான மொழி மாதிரியை மட்டுமல்லாமல் தொடக்க பாணி ஆப்டிகல் மாதிரி மற்றும் தனிப்பயன் டிகோடிங் வழிமுறையையும் உள்ளடக்கியது.

தொடக்க மாதிரி (புஜி மற்றும் பலர்., 2017)



தளவமைப்பு பகுப்பாய்வு

ஒழுங்கின் வாசிப்பைத் தீர்மானித்தல் மற்றும் தலைப்பு, தலைப்புகள் போன்றவற்றை வேறுபடுத்துதல்.

பரிசோதனை

முன்பு குறிப்பிட்டபடி, டெசராக்ட் கூகிள் ஸ்பான்சர் செய்கிறது. ஆசிரியர்கள் முடிவை டெசராக்டுடன் ஒப்பிடுவதற்கு இதுவும் ஒரு காரணம் என்று கருத்தப்படுகிறது. மாதிரி ஒப்பீட்டுக்கு எழுத்து பிழை விகிதம் (Character Error Rate (CER) ஏற்றுக்கொள்ளப்படுகிறது. இது திருத்த தூரமாக குறிப்பு நீளத்தால் வகுக்கப்பட்டு 100ஆல் அளவிடப்படுகிறது. குறைவானது சிறந்தது.

டெசராக்ட் Vs கூகிள் கிளவுட் விஷன் ஏபிஐ (வாக்கர் மற்றும் பலர். 2018)

Language	Books			Web		
	#Lines	N-CER [%]		#Lines	N-CER [%]	
		Tesseract	Google		Tesseract	Google
Arabic	946	14.0	4.8	4208	54.8	19.4
English	1000	1.0	0.6	4868	44.0	15.6
Hindi	1067	5.4	2.5	3726	49.3	20.6
Japanese	773	28.0	4.9	3256	57.5	17.1
Russian	864	1.7	1.2	3883	36.2	16.7

கூகிளின் ஒளிவழி எழுத்துணரி (OCR) மென்பொருள் 248+ மொழிகளுக்கு வேலை செய்கிறது

கூகிளின் ஆப்டிகல் கேரக்டர் ரெக்னிகிஷன் (OCR) மென்பொருள் இப்போது 248 க்கும் மேற்பட்ட உலக மொழிகளுக்கு (அனைத்து முக்கிய தெற்காசிய மொழிகளையும் உள்ளடக்கியது) செயல்படுகிறது. இது மிகவும் எளிமையானது மற்றும் பயன்படுத்த எளிதானது, மேலும் 90% க்கும் அதிகமான துல்லியத்துடன் பெரும்பாலான மொழிகளைக் கண்டறிய முடியும்.

தொழில்நுட்பம் படங்களிலிருந்து உரையை பிரித்தெடுக்கிறது, அச்சிடப்பட்ட உரையின் ஸ்கேன் மற்றும் கையெழுத்து கூட, அதாவது பழைய புத்தகங்கள், கையெழுத்துப் பிரதிகள் அல்லது படங்களிலிருந்து உரையை பிரித்தெடுக்க முடியும்.

கூகிளின் OCR அநேகமாக டெசராக்டின் சார்புகளை பயன்படுத்துகிறது, இது இலவச மென்பொருளாக வெளியிடப்பட்ட OCR இயந்திரம் அல்லது OCRopus, ஒரு இலவச ஆவண பகுப்பாய்வு மற்றும் ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன் (OCR) அமைப்பு முதன்மையாக கூகிள்

புத்தகங்களில் (Google Books) பயன்படுத்தப்படுகிறது. 1995-2006 ஆம் ஆண்டில் ஒரு சமூகத் திட்டமாக உருவாக்கப்பட்டது, பின்னர் கூகிள் கையகப்படுத்தியது, டெசராக்ட் மிகவும் துல்லியமான ஒளிவழி எழுத்துணரி இயந்திரங்களில் ஒன்றாகக் கருதப்படுகிறது மற்றும் 60 க்கும் மேற்பட்ட மொழிகளுக்கு வேலை செய்கிறது. மூல குறியீடு கிட்ஹப்பில் (GitHub) கிடைக்கிறது.

வெளியீட்டு உரையில் ஒளிவழி எழுத்துணரிக்குப் பிறகு தடித்த மற்றும் சாய்வு போன்ற விஷயங்களுக்கு எழுத்து வடிவமைப்பைப் பாதுகாப்பது குறித்த கூடுதல் விவரங்களை ஒளிவழி எழுத்துணரி திட்ட ஆதரவு பக்கம் வழங்குகிறது:

உங்கள் ஆவணத்தை செயலாக்கும்போது, தைரியமான மற்றும் சாய்வு உரை, எழுத்துரு அளவு மற்றும் வகை மற்றும் வரி முறிவுகள் போன்ற அடிப்படை உரை வடிவமைப்பைப் பாதுகாக்க முயற்சிக்கிறோம். இருப்பினும், இந்த கூறுகளைக் கண்டறிவது கடினம், நாம் எப்போதும் வெற்றிபெறாமல் போகலாம். புல்லட் மற்றும் எண்ணிடப்பட்ட பட்டியல்கள், அட்டவணைகள், உரை நெடுவரிசைகள் மற்றும் அடிக்குறிப்புகள் அல்லது இறுதி குறிப்புகள் போன்ற பிற உரை வடிவமைத்தல் மற்றும் கட்டமைக்கும் கூறுகள் தொலைந்து போக வாய்ப்புள்ளது.

தமிழ் மொழி விக்கிமீடியன் மற்றும் விக்கிமீடியா இந்தியாவின் திட்ட இயக்குனர் ரவிசங்கர் அய்யக்கண்ணு பேஸ்புக்கில் சோதனைக்குப் பின் இவ்வாறு கூறினார்: "மலையாளம் மற்றும் தமிழ் போன்ற சில மொழிகளுக்குத் தானியங்கிக் கத்தரிப்பு (auto cropping) போன்ற வடிவமைத்தல், உரையை பிரித்தல், படங்களை நிராகரித்தல் மற்றும் வண்ண பின்னணிகளை புறக்கணித்தல் இவற்றின் ஆதரவுடன் ஒளிவழி எழுத்துணரி கிட்டத்தட்ட 100% துல்லியத்துடன் செயல்படுகிறது." பங்களா, மலையாளம், கன்னடம், ஒடியா, தமிழ் மற்றும் தெலுங்கு ஆகிய இந்திய மொழிகளின் பூர்வீக பேச்சாளர்களும் ஓ.சி.ஆரைப் பரிசோதித்தபின் பேஸ்புக் பதிவில் கருத்துத் தெரிவித்தனர். இருப்பினும், குர்முகி (பஞ்சாபி எழுதப் பயன்படும்) போன்ற ஒரு சில ஸ்கிரிப்டுகளுக்கு,

ஓ.சி.ஆருக்குப் பிறகு வெளியீடு மிகவும் மோசமாக உள்ளது மற்றும் வெவ்வேறு ஸ்கிரிப்ட்களில் அபத்தமான உரையை விளைவிக்கிறது.

கூகிளின் ஒளிவழி எழுத்துணரியைப் பயன்படுத்தி ஸ்கேன் செய்யப்பட்ட படத்திலிருந்து ஒடியாவில் (இந்திய மொழி) உரையை மாற்றுவதற்கான பயிற்சியை சுபாஷிஷ் பனிகிராஹி வடிவமைத்துள்ளார்.

ஒட்டுமொத்தமாக, இது இன்னும் மின்னிலக்க மயமாக்கப்படாத பழைய நூல்களைக் கொண்ட மொழிகளுக்கு மிகப் பெரிய பாய்ச்சல். பல மொழிகளில் பழைய மற்றும் மதிப்புமிக்க உரையை இப்போது மின்னிலக்க மயமாக்கி விக்கிசோர்ஸ் (Wikisource) போன்ற தளங்களைப் பயன்படுத்தி இணையத்தில் பகிரலாம்.

தானியக்க அசைவூட்டிற்கான (automatic animation) சிறந்த எழுத்துணரி மென்பொருளை நாம் எவ்வாறு தேர்ந்தெடுப்போம்

வினையூக்கத்தின் புத்திசாலித்தனமான தானியக்க இயங்குதளத்தில் ஒளிவழி அறிதலுக்கான கூகிள் விஷனை (Google Vision) ஏன் பயன்படுத்துகிறோம்?

ஆவணங்களிலிருந்து தரவைப் பிரித்தெடுப்பது வினையூக்கத்தைப் (Catalytic's intelligent automation platform) பயன்படுத்தி கட்டப்பட்ட பல தானியங்கி செயல்முறைகளின் முக்கிய பகுதியாகும். கட்டமைக்கப்படாத தரவு மற்றும் ஆவணங்கள் சம்பந்தப்பட்ட நிகழ்வுகளுக்கு, ஒளிவழி எழுத்துணரியைப் (OCR) பயன்படுத்துவது சிறந்த தீர்வுகளில் ஒன்றாகும். ஆவண வாசிப்பு மற்றும் பிரித்தெடுப்பதற்கு எதைப் பயன்படுத்தவேண்டும் என்பதை வாடிக்கையாளர்கள் அடிக்கடி கேட்கிறார்கள்; எனவே கூகிள் விஷனைப் பயன்படுத்துவதற்கான முடிவுக்கு எவ்வாறு வந்தோம் என்பது குறித்த சில நுண்ணறிவை நாம் அறிந்துகொள்ள இயலும்.

புதிய அம்சத்தை உருவாக்கும்போது, நாம் தேர்வுசெய்யக்கூடிய நான்கு முக்கிய பாதைகள் உள்ளன:

- புதிதாக அதை நாமே தொடக்கத்திலிருந்து வளர்த்துக் கொள்ளவும்.
- முன்பே கட்டப்பட்ட தீர்வை வாங்கவும் அல்லது திறந்த மூல திட்டத்தைப் பயன்படுத்தவும், அதை நாமே ஹோஸ்ட் செய்யவும்
- ஒரு சுற்றுச்சூழல் அமைப்பு ஒருங்கிணைப்பை உருவாக்கி, ஒரு வினையூக்கச் செயலைச் செயல்படுத்த API ஐப் பயன்படுத்தவும்
- பல மூன்றாம் தரப்பு தயாரிப்புகளுடன் ஒருங்கிணைப்புகளை உருவாக்கவும் மற்றும் வாடிக்கையாளர்களைத் தங்கள் சொந்த கணக்கில் இணைக்க அனுமதிக்கவும்

ஒளிவழி எழுத்துணரி ஒருங்கிணைப்பை உருவாக்குதல்

ஒளிவழி எழுத்துணரியைப் பொறுத்தவரை, நான்கு பாதைகளை ஆராய்ந்த பிறகு, சுற்றுச்சூழல் அமைப்பு ஒருங்கிணைப்பை உருவாக்க நாம் தேர்வுசெய்தோம். கூகிள் OCR இன் சந்தைத் தலைவராக உள்ளார் என்ற முடிவுக்கு வந்தோம், மேலும் இந்த ஏழு காரணங்களுக்காக அந்த நிலையைத் தொடர்ந்து பராமரிக்க அதிக வாய்ப்புள்ளது:

1. ஒளிவழி எழுத்துணரி மற்றும் கூகிளில் சந்தை தலைவர்களை நாங்கள் சோதித்து ஒப்பிட்டுப் பார்க்கலாம். ஒளிவழி எழுத்துணரிக்கான துல்லியமான துல்லிய எண்களை வழங்குவது கடினம், ஏனென்றால் ஒரு ஆவணத்தின் வடிவம் மற்றும் தரத்தின் அடிப்படையில் முடிவுகள் பெருமளவில் வேறுபடுகின்றன. ஆனால் விலைப்பட்டியல் மற்றும் ஒப்பந்தங்கள் மற்றும் PDFகள் மற்றும் JPEG கள் போன்ற கோப்பு வகைகளின் பகுப்பாய்வு வகைகளின் பகுப்பாய்வின் அடிப்படையில், குறைந்த தெளிவுத்திறன் கொண்ட கோப்புகளுக்குக் கூட கூகிள் தொடர்ந்து சிறந்த எழுத்து அங்கீகாரத்தை வழங்குவதைக் கண்டறிய இயலும்.

2. கூகிள் அதிக அளவு பாதுகாப்பைக் கொண்டுள்ளது. HIPAA மற்றும் SOC2 வகை 2 இணக்கத்துடன் ஒழிவழி எழுத்துணருலுக்கான நம் பாதுகாப்பு தரங்களை நிறுவனம் பூர்த்தி செய்கிறது அல்லது மீறுகிறது.

3. கூகிள் ஒளிவழி எழுத்துணரி மிகவும் நம்பகமான மற்றும் செயல்திறன் மிக்கது. இது பெரிய ஆவணங்களை விரைவாக செயலாக்க முடியும் மற்றும் கூகிள் செயலிழப்புகள் மிகவும் அரிதானவை, அல்லது அவை நிகழும்போது, மிகக் குறுகியவை.

4. கூகிள் தனது ஒளிவழி எழுத்துணரி தொழில்நுட்பத்தை மேம்படுத்த ஆராய்ச்சியில் தொடர்ந்து முதலீடு செய்கிறது. முன்னணி திறந்த மூல ஒளிவழி எழுத்துணரி தயாரிப்பான டெசராக்ட்டின் முக்கிய ஸ்பான்சர் இந்நிறுவனம்.

5. கூகிள் ஒளிவழி எழுத்துணரி மற்றும் கணினி பார்வையில் பல தசாப்தங்களாக அனுபவம் கொண்டுள்ளது. இது Android கேமரா, கூகிள் புகைப்படங்கள், வேமோ, படத் தேடல் மற்றும் வீதிக் காட்சி போன்ற பல தயாரிப்புகளின் அடிப்படை பகுதியாகும்.

6. கூகிளின் ஏபிஐக்கள் (Google's APIs) நன்கு வடிவமைக்கப்பட்டுள்ளன, இது கூகிள் விஷனில் சேர்க்கப்பட்ட புதிய அம்சங்களையும் செயல்பாடுகளையும் பின்பற்றுவதற்கான வினையூக்கத்திற்கான குறுகிய வளர்ச்சி நேரத்திற்கு வழிவகுக்கிறது.

7. புதியவற்றைச் சேர்க்கும்போது, கிடைக்கக்கூடிய அம்சங்களின் முடிவுகளின் தரத்தை கூகிள் தொடர்ந்து மேம்படுத்துகிறது. எடுத்துக்காட்டாக, கூகிள் டிசம்பர் 2018 இல் கையெழுத்து ஒளிவழி எழுத்துணரியை வெளியிட்டபோது, வினையூக்கி (Catalytic) உடனடியாக அந்த அம்சத்தையும் சேர்த்தது மற்றும் ஆதரித்தது.

கூகிள் பார்வை ஒப்பீடுகள்

கூகிள் பார்வை பிற ஒளிவழி எழுத்துணரி வழங்குநர்களுடன் ஒப்பிடுவதற்கான சில ஆதாரங்கள் கீழே தரப்பட்டுள்ளன:

• ஆய்வு கண்டுபிடிப்புகள் கூகிள் விஷன் சிறந்த பட உணர்தல் அமைப்பு | CIO டைவ் (CIO Dive)

• சிறந்த பட உரை உணர்தல் API களை ஒப்பிடுவது | தரவு துருக்கியர்கள் (Data Turks)

• பட உணர்தல் துல்லியம் ஆய்வு | சரியான டிஜிட்டல் (Perficient Digital)

• சிறந்த ஒளிவழி எழுத்துணரி கருவிக்கான நம் தேடல், நாம் கண்டுபிடித்தது | ஆதாரம்: ஒரு ஓபன்நியூஸ் திட்டம் (OpenNews project)

இந்த அனைத்து காரணிகளையும் அடிப்படையாகக் கொண்டு, கூகிள் விஷன் ஏபிஐ உடன் சுற்றுச்சூழல் அமைப்பு ஒருங்கிணைப்பை உருவாக்குவது என்பது தன்னியக்கத்திற்கான ஒளிவழி எழுத்துணரியின் பல பயன்பாடுகளுக்கு சக்தி அளிப்பதற்கான சிறந்த விருப்பத்தை அளிக்கிறது.

ஒளிவழி எழுத்துணரி என்பது வேகமாக வளர்ந்து வரும் இடமாகும், மேலும் அமேசான் டெக்ஸ்ட்ராக்ட் (Amazon Textract) மற்றும் மைக்ரோசாஃப்ட் கம்ப்யூட்டர் விஷன் (Microsoft Computer Vision) போன்ற வணிக கிளவுட் வரவுகள் மற்றும் பல புதிய தொடக்கங்களால் (new startups) அற்புதமான முன்னேற்றங்கள் உள்ளன. இது நாம் தொடர்ந்து கண்காணித்து வரும் ஒரு பகுதியாகும், எனவே கிடைக்கக்கூடிய சிறந்த ஒளிவழி எழுத்துணரி தொழில்நுட்பத்தைப் பயன்படுத்தி கட்டமைக்கப்படாத தரவுகளிலிருந்து தகவல்களைப் பெறுவதற்கான வினையூக்கத்தின் (Catalytic's) திறனை தொடர்ந்து மேம்படுத்தலாம்.

வினையூக்கியுடன் ஒளிவழி எழுத்துணரியைப் பயன்படுத்துதல்

கிளவுட் இயங்குதளத்துடன் இணைக்க வினையூக்கி தொடர்ந்து ஆராய்ச்சி செய்யும் கருவிகளில் ஒன்றாகும் ஒளிவழி எழுத்துணரி, எனவே ஆட்டோமேஷன்களை உருவாக்குவதற்கான எளிதான வழியை நாம் பெற முடியும். இயற்கையான மொழி செயலாக்கம் (Natural Language Processing (NLP/என்.எல்.பி), தரவுத்தள நுழைவு, மின்னஞ்சல் வார்ப்புரு நிரப்புதல், AI-இயங்கும் உணர்வு பகுப்பாய்வு (AI-powered sentiment analysis) அல்லது ஒரு முடிவுக்கு முடிவு செயல்முறையில் அடுத்த கட்டமாக முடிவெடுப்பது ஆகியவற்றை உள்ளடக்கிய முழு தானியங்கி

செயல்முறை முழுவதும் கூகிள் விஷன் ஓ.சி.ஆருடன் பிரித்தெடுக்கப்பட்ட தகவல்களைப் பயன்படுத்த வினையூக்கி தளம் உங்களை அனுமதிக்கிறது.

மின்னிலக்க உருமாற்றத்திற்கான உங்கள் எளிதான பாதைக்கான கட்டடத் தொகுதிகளை வினையூக்கியின் தளம் வழங்குகிறது. ஒளிவழி எழுத்துணரி பலவற்றில் ஒன்றாகும். புத்திசாலித்தனமான தானியக்கத்திற்காக சிறந்த ஒளிவழி எழுத்துணரியை எவ்வாறு பயன்படுத்துவது என்பதைப் பார்க்க எங்கள் நிபுணர்களுடன் ஒரு டெமோவைத் திட்டமிடுங்கள்.

மின்னிலக்க உருமாற்ற முயற்சிகள் நிறுத்தப்படுவதற்கான முதல் மூன்று காரணங்களுக்காக முழுக்குங்கள், மற்றும் உறுதியான பயன்பாட்டு நிகழ்வுகள் இருப்பினும், குர்முகி (பஞ்சாபி எழுதப் பயன்படும்) போன்ற ஒரு சில ஸ்கிரிப்டுகளுக்கு, ஓ.சி.ஆருக்குப் பிறகு வெளியீடு மிகவும் மோசமாக உள்ளது மற்றும் வெவ்வேறு ஸ்கிரிப்ட்களில் அபத்தமான உரையை விளைவிக்கிறது.

ஒட்டுமொத்தமாக, இது இன்னும் டிஜிட்டல் மயமாக்கப்படாத பழைய நூல்களைக் கொண்ட மொழிகளுக்கு மிகப் பெரிய பாய்ச்சல். பல மொழிகளில் பழைய மற்றும் மதிப்புமிக்க உரையை இப்போது டிஜிட்டல் மயமாக்கி விக்கிசோர்ஸ் போன்ற தளங்களைப் பயன்படுத்தி இணையத்தில் பகிரலாம்.

வினையூக்கி பற்றி (About Catalytic)

வினையூக்கி என்பது ஒரு குறியீடு இல்லாத பணிப்பாய்வு ஆட்டோமேஷன் தளமாகும் (no-code workflow automation platform), இது வணிக மக்களை குடிமக்கள் உருவாக்குநர்களாக மாற்றுவதற்கும், டிஜிட்டல் மயமாக்கப்பட்ட, தானியங்கு மற்றும் செயற்கையறிவு-செயல்படுத்தப்பட்ட பணிப்பாய்வுகளை (AI-enabled workflows) விரைவாக உருவாக்குவதற்கும் உதவுகிறது. வினையூக்கி மூலம், நிறுவனங்கள் அதிக செயல்திறன் மற்றும் குறைந்த உராய்வுடன் வேகமாக செயல்பட முடியும்.

4. தமிழ் கையெழுதுப் படிவத்திற்கான எழுத்துணரித் தொழில்நுட்பம் (Hand written OCR in Tamil)

ஒளிவழி எழுத்துணரி நுட்பங்களைப் பயன்படுத்தி தமிழ் கையால் எழுதப்பட்ட எழுத்து உணர்தல் குறித்த ஒரு ஆய்வு

இன்றைய வேகமாக வளர்ந்து வரும் தொழில்நுட்பத்தில், டிஜிட்டல் உணர்தல் பரந்த பங்கைக் கொண்டுள்ளன; மேலும் ஒளிவழி எழுத்துணரி நுட்பங்களில் ஆராய்ச்சி செய்ய அதிக வாய்ப்பை வழங்குகின்றன. மற்ற மேற்கத்திய மொழி ஸ்கிரிப்டுகளுடன் ஒப்பிடும்போது தமிழ் கையால் எழுதப்பட்ட ஸ்கிரிப்டுகளின் உணர்தல் சிக்கலானது. இருப்பினும், பல ஆராய்ச்சியாளர்கள் ஆஃப்லைன் தமிழ் எழுத்து உணர்தலுகும் நிகழ்நேர தீர்வுகளை வழங்கியுள்ளனர். ஆஃப்லைன் தமிழ் கையால் எழுதப்பட்ட ஆவணங்கள் உணர்தல் இன்னும் ஆராய்ச்சியாளர்களுக்கு பல ஊக்கமளிக்கும் சவால்களை வழங்குகிறது. தற்போதைய ஆராய்ச்சி தமிழ் கையால் எழுதப்பட்ட ஆவணங்களை உணர்வதில் பல தீர்வுகளை வழங்குகிறது, ஆனால் நியாயமான துல்லியம் மற்றும் செயல்திறன் அடையப்படவில்லை. \ தமிழ் கையால் எழுதப்பட்ட எழுத்து உணர்தல் தொடர்பான பல்வேறு அணுகுமுறைகள் மற்றும் சவால்களை பகுப்பாய்வு செய்கிறது. டிஜிட்டல் உள்ளடக்கம் தோன்றியவுடன், உயர் செயல்திறன் கொண்ட ஒளிவழி எழுத்துணரி இயந்திரத்தின் வளர்ச்சிக்கான தேவை அவசியமாகிவிட்டது. ஒளிவழி எழுத்துணரி ஆராய்ச்சி பணிகள் பல ஆராய்ச்சியாளர்களால் மேற்கொள்ளப்பட்டுள்ளன, அவை உயர் செயல்திறன் கொண்ட ஒளிவழி எழுத்துணரி இயந்திரத்தை உருவாக்குவதை நோக்கமாகக் கொண்டுள்ளன. ஒரு ஒளிவழி எழுத்துணரிக்குப் பின்னால் உள்ள யோசனை என்னவென்றால், ஒரு ஆவணப் படத்தை பக்கத்தை வரி கூறுகளாகப் பிரிப்பதன் மூலமும், மேலும் சொற்களாகப் பிரிப்பதன் மூலமும் பின்னர் எழுத்துக்களாகப் பிரிப்பதன் மூலமும் அடையாளம் காணவும் பகுப்பாய்வு செய்யவும். சாத்தியமான எழுத்துக்களைக் கணிக்க இந்த எழுத்துக்கள் பட வடிவங்களுடன் ஒப்பிடப்படுகின்றன. எழுத்துக்களை உணர்வது அச்சிடப்பட்ட

ஆவணங்களிலிருந்து அல்லது கையால் எழுதப்பட்ட ஆவணங்களிலிருந்து செய்யப்படலாம். கையால் எழுதப்பட்ட ஆவண உணர்தல் ஆஃப்லைனில் அல்லது ஆன்லைனில் செய்யப்படலாம். ஆன்லைனை விட ஆஃப்லைன் எழுத்து உணர்தல் மிகவும் சிக்கலானது. குறிப்பாக, தமிழ் கையால் எழுதப்பட்ட ஒளிவழி எழுத்துணரி மற்ற தொடர்புடைய படைப்புகளை விட மிகவும் சிக்கலானது. ஏனென்றால், தமிழ் எழுத்துக்களில் அதிக கோணங்களும் மாற்றிகளும் உள்ளன. கூடுதலாக, தமிழ் ஸ்கிரிப்டில் அதிக எண்ணிக்கையிலான எழுத்துத் தொகுப்புகள் உள்ளன. மொத்தம் 247 எழுத்துக்கள்; 216 கூட்டு எழுத்துக்கள், 18 மெய், 12 உயிரெழுத்துகள் மற்றும் ஒரு சிறப்பு எழுத்து ஆகியவற்றைக் கொண்டுள்ளது. அங்கீகார செயல்பாட்டின் போது எதிர்கொள்ளும் சவால்கள் எழுத்துக்களில் உள்ள வளைவுகள், கோடுகள் மற்றும் துளைகளின் எண்ணிக்கை, நெகிழ் எழுத்துக்கள், மாறுபட்ட எழுத்து நடைகள் போன்றவை. எழுத்துக்குறி உணர்தலில் ஈடுபடும் படிகள் முன் செயலாக்கம், பிரிவு, அம்சம் பிரித்தெடுத்தல் மற்றும் வகைப்பாடு ஆகியவற்றை உள்ளடக்கியது. புள்ளிவிவர, கட்டமைப்பு மற்றும் கலப்பின என்ற மூன்று வகையான அம்சங்களை அங்கு பகுப்பாய்வு செய்யலாம்.

1.**முன் செயலாக்கம்**

எழுத்து உணர்தலைச் செய்வதற்கு முன் ஏராளமான பணிகள் முடிக்கப்பட உள்ளன. கையால் எழுதப்பட்ட ஆவணம் ஸ்கேன் செய்யப்பட்டு செயலாக்கத்திற்கு ஏற்ற வடிவமாக மாற்றப்பட வேண்டும். முன் செயலாக்கம் ஆவணப் படத்தை சுத்தம் செய்வதற்கும் உணர்தல் செயல்முறையை துல்லியமாக எடுத்துச் செல்வதற்கும் பொருத்தமான சில வகையான துணை செயல்முறைகளைக் கொண்டுள்ளது. முன் செயலாக்கத்தில் ஈடுபடும் துணை செயல்முறைகள் கீழே விளக்கப்பட்டுள்ளன:

- இருமையாக்கம் (Binarization)
- சத்தம் குறைப்பு (Noise reduction)

• இயல்பாக்குதல் (Normalization)

• வளைவு திருத்தம், மெல்லிய மற்றும் சாய்ந்த நீக்கம் (Skew correction, thinning and slant removal)

1) இருமையாக்கம் (Binarization)

இருமையாக்கம் (Binarization) என்பது ஒரு சாம்பல் அளவுப் படத்தை (gray scale image) தொடக்கநிலை (thresholding) மூலம் கருப்பு மற்றும் வெள்ளை படமாக மாற்றும் ஒரு முறையாகும். மற்றொரு அணுகுமுறை, இருமையாக்கம் செய்யப்பட்ட படத்தை (binarized image) தானாகப் பெற ஹிஸ்டோகிராம் (histogram) அடிப்படையிலான த்ரெஷோல்டிங்கைச் செய்ய ஓட்சுவின் முறை பயன்படுத்தப்படலாம். மல்டி ஓஸ்டு முறை (Otsu's method) எனப்படும் பல நிலை தொடக்கநிலைகளுக்கு ஓட்சுவின் முறை நீட்டிக்கப்பட்டுள்ளது. பொதுவாக, பெரும்பாலான ஆராய்ச்சியாளர்கள் பின்னணி படத்திலிருந்து முன் படத்தை பிரித்தெடுக்க தொடக்கநிலை கருத்துருக்களைப் பயன்படுத்துகின்றனர். இந்த முறையில், இரண்டு முன்புற சாம்பல் குறியீடு படங்களுக்கு இடையில் எந்த மதிப்பையும் எடுத்துக்கொள்வதன் மூலம் தொடக்கநிலை மதிப்பு சரி செய்யப்படுகிறது. சாம்பல் அளவிலான படத்தை இரண்டு தொனி படமாக மாற்ற ஹிஸ்டோகிராம் அடிப்படையிலான வாசல் அணுகுமுறை பயன்படுத்தப்படலாம். இதற்கு நேர்மாறாக, படத்தின் இடம்சார்ந்த சாம்பல் மதிப்பு வேறுபாட்டை அடையாளம் காண தகவமைப்பு பைனரைசேஷன் முறையும் (Adaptive Binarization method) பயன்படுத்தப்படலாம். குறைந்த தரமான ஆவணங்களிலிருந்து உரை தகவல்களைப் பெற இது உதவும். இருநிலை உலகமய இருமையாக்க நுட்பம் (Two-Level Global Binarization Technique) என்ற மற்றொரு அணுகுமுறை உலகளாவிய த்ரெஷோல்டிங் நுட்பத்தைப் பயன்படுத்தி வெளியீட்டை உருப்படுத்தம் செய்கிறது.

2) சத்தம் அகற்றுதல்

மின்னிலக்கப் படங்கள் பல வகையான சத்தங்களுக்கு ஆளாகின்றன. ஆவணப் படத்தில் சத்தம் மோசமாக நகலெடுக்கப்பட்ட பக்கங்களால் ஏற்படுகிறது. சராசரி வடிகட்டுதல் (Median filters), வீனர் வடிகட்டுதல் முறை (Wiener Filtering method) மற்றும் சத்தத்தை அகற்ற உருவவியல் செயல்பாடுகள் செய்யப்படலாம். எழுத்துக்குறி படத்தின் தீவிரத்தை மாற்றுவதற்கு சராசரி வடிப்பான்கள் பயன்படுத்தப்படுகின்றன, காஸியன் வடிப்பான்கள் படத்தை மென்மையாக்க பயன்படுத்தலாம்.

3) இயல்பாக்கம் (Normalization)

இயல்பாக்கம் என்பது ஒரு சீரற்ற அளவிலான படத்தை ஒரு நிலையான அளவாக மாற்றும் செயல்முறையாகும். படத்திலிருந்து ஒற்றை கட்டமைப்பு உறுப்பு பெற RoiExtraction முறை பயன்படுத்தப்படுகிறது. Bicubic interpolation, நேரியல் அளவு இயல்பாக்கம் மற்றும் ஜாவா பட வகுப்பு இயல்பாக்குதல் நுட்பங்கள் நிலையான அளவிலான படங்களுக்கு பயன்படுத்தப்படலாம். பல படைப்புகளில், உள்ளீட்டுப் படம் 50 x 50 அளவிற்கு இயல்பாக்கப்படுகிறது, ஒவ்வொரு கையால் எழுதப்பட்ட படத்தின் எல்லை பெட்டியைக் கண்டறிந்த பின்னர் செயலாக்கத்தை எழுப்புகிறது.

4) வளைவு திருத்தம், மெல்லிதல் மற்றும் சாய்வு நீக்கம் (Skew correction, Thinning and Slant removal)

மெல்லிய ஒரு முன் செயல்முறை, இது ஒற்றை பிக்சல் அகல படத்தைக் கையால் எழுதப்பட்ட எழுத்தை எளிதில் அடையாளம் காணும். பட எழுத்துக்களின் பிக்சல் அளவிலான நேரியல் பிரதிநிதித்துவங்களை மட்டுமே விட்டுவிட்டு இது மீண்டும் மீண்டும் பயன்படுத்தப்படுகிறது. கபோர் வடிப்பான்களுடன் (Gabor filters) கூடிய விண்டோஸ் உரைத் தொகுதியின் ஒட்டுமொத்த அளவிடல் தயாரிப்பு (Cumulative scalar product) (CSP/சிஎஸ்பி) மெல்லிய நோக்கத்திற்காகப் பயன்படுத்தப்படுகிறது. உருபனியல் அடிப்படையிலான மெல்லிய

வழிமுறை மற்றும் பிற மெல்லிய வழிமுறைகள் சிறந்த குறியீட்டு பிரதிநிதித்துவத்திற்கும் எழுத்து உருவங்களை மெல்லியதாகவும் பயன்படுத்தப்படுகின்றன. எலும்புக்கூடாக்கம் (Skeletonization) என்பது ஒரு வடிவத்தை பொதுவான வடிவத்தை பாதிக்காமல் முடிந்தவரை பல பிக்சல்களுக்கு சிதறடிக்கும் செயல்முறையாகும். படங்களிலிருந்து தேவையற்ற பிக்சல்களை அகற்றுவதற்கும் ஹில்டிட்சின் வழிமுறை பயன்படுத்தப்படுகிறது (எலும்புக்கூடாக்கம்/skeletonization). ஆவண ஸ்கேனிங்கின் போது உள்வரும் ஆவண படத்தில் வளைவு தவிர்க்க முடியாமல் அறிமுகப்படுத்தப்படுகிறது. இயல்பாக்கம், ஃபோரியர் ஸ்பெக்ட்ரம் நுட்பங்கள் (Fourier Spectrum techniques) சாய்வு, கோண பக்கவாதம், அகலம் மற்றும் செங்குத்து அளவிடுதல் ஆகியவற்றை சரிசெய்ய பயன்படுத்தப்படுகின்றன.

2. பிரித்தல் (Segmentation)

பிரித்தல் என்பது ஒரு செயல்முறையாகும், இது ஆவணப் படங்களை கோடுகள், சொற்கள் மற்றும் எழுத்துக்களாகப் பிரிக்கப் பயன்படுகிறது. வகை-எழுதப்பட்ட ஆவணங்களை (type-written documents) விடக் கையால் எழுதப்பட்ட ஆவணங்களின் பிரித்தல் மிகவும் சிக்கலானது. ஹிஸ்டோகிராம் சுயவிவரங்கள் (Histogram profiles) மற்றும் இணைக்கப்பட்ட கூறு பகுப்பாய்வு ஆகியவை வரிப் பிரித்தலுக்குப் (line segmentation) பயன்படுத்தப்படுகின்றன. பிரித்தல் செயல்பாட்டில், பத்திகளை அடையாளம் காண பத்தி இடம் சரிபார்க்கப்படும். கிடைமட்ட கோடுகளின் அகலத்தைக் கண்டறிய படத்தின் ஹிஸ்டோகிராம் பயன்படுத்தப்படும். சொல் பிரித்தலுக்கு இடஞ்சார்ந்த இடத்தைக் கண்டறியும் நுட்பம் பயன்படுத்தப்பட்டுள்ளது. சொற்களின் இரு அகலத்தையும் கண்டறிவதற்கும் படத்தை கிளிஃபாக மாற்றுவதற்கும் ஹிஸ்டோகிராம் முறைகள் பயன்படுத்தப்பட்டுள்ளன. எல்லை என்ற வார்த்தையை அடையாளம் காண வரிகளுக்குள் இடைவெளியைக் கண்டறிய செங்குத்து ஹிஸ்டோகிராம் சுயவிவர முறை பயன்படுத்தப்படும். படத்திலிருந்து தனிப்பட்ட எழுத்துக்களைப் பெற பிராந்திய ஆய்வு

அல்காரிதம் பயன்படுத்தப்பட்டும். ப்ரொஜெக்ஷன் அடிப்படையிலான, ஸ்மியரிங், குழுமம், ஹஃப்-அடிப்படையிலான, வரைபட அடிப்படையிலான மற்றும் வெட்டு உரை குறைத்தல் (சி.டி.எம்) என வெவ்வேறு நிலைகளின் அடிப்படையில் பிரித்தல் வகைப்படுத்தப்படும். மாற்றியமைக்கப்பட்ட குறுக்கு எண்ணும் நுட்பம், ஹிஸ்டோகிராம் சுயவிவரம் மற்றும் இணைக்கப்பட்ட கூறு பகுப்பாய்வு ஆகியவை வரி எழுத்துப் பிரிவு பிரித்தல் சிக்கலைக் கையாள ஆய்வில் காணப்படுகின்றன.

3. பண்புக்கூறு பிரித்தெடுத்தல்

பண்புக்கூறு பிரித்தெடுக்கும் நுட்பங்களை புள்ளிவிவர பண்புக்கூறுகள், கட்டமைப்பு பண்புக்கூறுகள் மற்றும் கலப்பின பண்புக்கூறுகள் என மூன்று வகுப்புகளாகப் பிரிக்கலாம். ஒரு புள்ளிவிவர நுட்பம் பண்புக்கூறுகளைப் பிரித்தெடுப்பதற்கான அளவு அளவீடுகளைப் பயன்படுத்துகிறது, அதேசமயம் கட்டமைப்பு நுட்பங்கள் பண்புக்கூறுகள் பிரித்தெடுப்பதற்கான தரமான அளவீடுகளைப் பயன்படுத்துகின்றன. கலப்பின அணுகுமுறையில், இந்த இரண்டு நுட்பங்களும் ஒன்றிணைக்கப்பட்டு உணர்தலுக்கு பயன்படுத்தப்படுகின்றன.

1) **கட்டமைப்பு நுட்பம் (Structural Technique)**

எழுத்துக்குறி படத்தை இடம்சார் பண்புக்கூறுகளின் தொகுப்பாக மாற்ற, அளவு மாறா பண்புக்கூறு மாற்றம் (Scale Invariant feature transform (SIFE) பயன்படுத்தப்படுகிறது. இந்த அணுகுமுறையைப் பயன்படுத்தி, SIFE பண்புக்கூறுகளின் 128 பரிமாணங்கள் (சுவாரஸ்யமான புள்ளிகள்) எழுத்து படத்திலிருந்து அடையாளம் காணப்படுகின்றன. ஒரு படம் இரண்டு தொனி படமாக (two tone image) மாற்றப்பட்டு, பின்னர் சட்டமாக மாற்றப்படுகிறது. சட்டத்திலிருந்து பெறப்பட்ட பிரேம் புள்ளி திசையன்களைக் (vectors) கொண்டிருக்கும். இயல்பாக்கப்பட்ட பண்புக்கூறு திசையன் (Normalized Feature Vector NFV) திசையனிடமிருந்து அதாவது கோடுகள் மற்றும் வளைவுகளிலிருந்து முன்மாதிரி பெறுகிறது.

2) **புள்ளிவிவர நுட்பம் (Statistical Technique)**

மண்டல அடிப்படையிலான முறையில் (Zone based method), இயல்பாக்கப்பட்ட எழுத்துக்கள் இடைப்பின்னாத மண்டலங்களாக (non interleaving zones) பிரிக்கப்படுகின்றன. ஒவ்வொரு மண்டலத்திற்கும் பிக்சல் அடர்த்தி கணக்கிடப்படுகிறது. இது பண்புக்கூறுகளைக் குறிக்கப் பயன்படுகிறது. குறியீட்டுப் பைனரி மாறுபாடு (encoding Binary variation approach) அணுகுமுறையைப் பயன்படுத்தி எழுத்து பிக்சல்களின் உயரமும் அகலமும் கணக்கிடப்படுகின்றன. வரிசை மற்றும் அகலத்தின் உயர் மட்டத்தை அடைந்ததும், செயல்முறை நிறுத்தப்படும். அணுகுமுறை மற்றும் அதிலிருந்து பிரித்தெடுக்கப்பட்ட பண்புக்கூறுகளின்படி ஒரு பைனரி கொடி (binary flag) அமைக்கப்படுகிறது. கபார் சேனல் முறையைப் (Gabor channel method) பயன்படுத்தி படங்கள் சம அளவிலான ஒன்பது அல்லாத ஒன்றுடன் ஒன்று தொகுதிகளாக பிரிக்கப்படுகின்றன. இது ஒவ்வொரு சேனலிலும் செல்லும் தொகுதிகளுக்கு 24 பதில்களை வழங்குகிறது. சராசரி மற்றும் நிலையான வழித்தோன்றல்கள் கணக்கிடப்பட்டு பண்புக்கூறுகளாகப் பயன்படுத்தப்படுகின்றன.

பிலினியர் இன்டர்போலேஷன் டெக்னிக் பயன்படுத்தி (Bilinear Interpolation Technique) அனைத்து படங்களும் ஒரே உயரம் மற்றும் அகலத்திற்கு அளவிடப்படுகின்றன. தேவையற்ற பகுதிகள் சோபல் விளிம்பு கண்டறிதல் வழிமுறையைப் (Sobel edge detection algorithm) பயன்படுத்தி சரி செய்யப்படுகின்றன. RoiExtraction அணுகுமுறை மூடிய படத்தின் விளைவாக உருவத்தின் உருபனியல் மூடுதலைப் பயன்படுத்துகிறது. இந்த அணுகுமுறை வரம்பிலிருந்து இடதுபுறம், வலதுபுறம், மேல் மற்றும் கீழ்மட்டத் தொகுதியைத் தடுக்கிறது. பண்புக்கூறுகள் இயல்பாக்கப்பட்ட x-y ஆயத்தொகுப்புகள்; முதல் மற்றும் இரண்டாவது ஆக்கங்கள், வளைவு, பண்புக்கூறு, சுருள் மற்றும் மெல்லிய தன்மை ஆகியவை டைம் டொமைன் பண்புக்கூறுகளிலிருந்து பெறப்படுகின்றன. ஸ்லைடிங் ஹாமிங் சாளரத்தைப் (sliding hamming window) பயன்படுத்தி கோட்டுடன் அதிர்வெண் களத்திற்கான பண்புக்கூறுகள்

கணக்கிடப்படுகின்றன. குறைந்த குணகத்தின் உண்மையான மற்றும் கற்பனையான பகுதி பண்புக்கூறு திசையனுடன் சேர்க்கப்படுகிறது. இறுதி புள்ளி, கிளைப் புள்ளி, துளைகள், நீளம், வடிவம் மற்றும் தனிப்பட்ட கோட்டின் வளைவு போன்ற கட்டமைப்பு பண்புக்கூறுகள் ஆக்டல் வரைபடத்தைப் (octal graph) பயன்படுத்தி பெறப்படுகின்றன. படங்களின் எல்லைகளைக் கண்டறிய "எட்டு அண்டை" சரிசெய்தல் முறை ("eight-neighbour" adjustment method) பயன்படுத்தப்படுகிறது. பைனரி படத்தின் எல்லையைக் கண்டுபிடிக்கும் வரை அணுகுமுறை ஸ்கேன் செய்கிறது. பின்னர், ஃபோரியர் டிஸ்கிரிப்டர் குணகத்தைக் கண்டுபிடித்து மொத்த எல்லைகளின் எண்ணிக்கையைப் பெற பயன்படுத்தப்படுகிறது. இந்த வகை மாறாத விளக்கங்கள் மேலும் வகைப்படுத்தலுக்கு ஒரு நரம்பியல் நெட்வொர்க்கின் உள்ளீடாக வழங்கப்படுகின்றன.

3) கலப்பின நுட்பம்

கிடைமட்ட மற்றும் செங்குத்து கோடுகளைக் கண்டறிய ஹஃப் டிரான்ஸ்ஃபார்ம் (Hough Transform) பயன்படுத்தப்படுகிது. அவை மற்றொரு எளிய வழிமுறையைப் பயன்படுத்தி கிளை மற்றும் நிலையை ஆய்வு செய்து வருகின்றன. சாய்ந்த மற்றும் துண்டு போன்ற பண்புக்கூறுகளைப் பிரித்தெடுக்க பிலினியர் இடைக்கணிப்பு (Bilinear interpolation) என்ற மற்றொரு நுட்பம் பயன்படுத்தப்படுகிறது. ஒரு பண்புக்கூறு பிரித்தெடுத்தல் நுட்பம் மண்டல அடிப்படையிலான கலப்பின அணுகுமுறை, இது மண்டல சென்ட்ராய்டு (zone centroid) மற்றும் பட சென்ட்ராய்டு (Image centroid) அடிப்படையிலான தூர மெட்ரிக் பண்புக்கூறுகளைப் பிரித்தெடுக்கப் பயன்படுகிறது.

4) வகைப்படுத்தல்

பிரித்தெடுக்கப்பட்ட பண்புக்கூறுகள் வகைப்பாடு செயல்முறைக்கான உள்ளீடாக வழங்கப்படுகின்றன. பண்புக்கூறு பிரித்தெடுக்கும் அணுகுமுறைகளிலிருந்து (feature extraction approaches) பிரித்தெடுக்கப்பட்ட ஒரு முக்கிய விஷயங்களின் பை (பாக்-ஆஃப்-கீபாயிண்ஸ்/bag-

of-keypoints)) வகைப்படுத்தலுக்குப் பயன்படுத்தப்படுகின்றன. கே-அருகிலுள்ள அண்டையர் அணுகுமுறை (K-Nearest Neighbour approach), தெளிவில்லாத அமைப்பு, நரம்பியல் நெட்வொர்க், பாகுபாடு காண்பிக்கும் வகைப்படுத்தி, மேற்பார்வை செய்யப்படாத வகைப்படுத்தி மற்றும் பல போன்ற தற்போதைய அமைப்புகளில் உள்ள எழுத்து பண்புகூறுகளை வகைப்படுத்த சில அணுகுமுறைகள் பயன்படுத்தப்படுகின்றன.

கே-அருகிலுள்ள அண்டையர் அணுகுமுறை (K-Nearest Neighbour approach) எழுத்துக்குறி தொகுப்புகளையும் சிறந்த துல்லியத்தையும் உணர வகைப்படுத்தியாகப் பயன்படுத்தப்படுகிறது. உணரலின் முடிவு ஹிஸ்டோகிராம் சமன்பாட்டைப் (Histogram Equalization) பயன்படுத்தி பெறப்படுகிறது. சுய ஒழுங்குமுறை வரைபட அணுகுமுறையில் (Self organizing map approach) ஒவ்வொரு முனையின் எடையும் யூக்ளிடியன் தூர முறையைப் (Euclidean distance method) பயன்படுத்தி கணக்கிடப்படுகிறது. பின்னர், உள்ளீட்டு எழுத்துக்கள் SOM மூலம் கணக்கிடப்பட்ட அனைத்து திசையன்களுடன் ஒப்பிடப்படுகின்றன. அவை பொருந்தினால், வெளியீடு அங்கீகரிக்கப்பட்ட திசையனாக வழங்கப்படும்.

இது சிறந்த பொருத்தம் அலகு (Best Matching Unit (BMU/பி.எம்.யூ) என்று அழைக்கப்படுகிறது. இந்த அணுகுமுறை வகைப்படுத்தத் தவறினால் உலகளாவிய அம்சங்கள் பயன்படுத்தப்படுகின்றன. இரண்டு மறைக்கப்பட்ட அடுக்கு அணுகுமுறைக்கான மல்டிலேயர் பார்வையில் (Multilayer perception), 1 மற்றும் -1க்கும் இடையில் ஒரு சீரற்ற எண்ணுடன் பொருத்தப்பட்ட அனைத்து எடைகளையும் கொண்டு ஒரு நரம்பியல் பிணையம் வடிவமைக்கப்பட்டுள்ளது. சிறந்த அங்கீகார முடிவைக் கண்டறிய எழுத்துக்கள் பிணைய வெளியீட்டோடு ஒப்பிடப்படுகின்றன. சிறந்த செயல்திறனுக்காக இரண்டு மறைக்கப்பட்ட அடுக்குகள் பயன்படுத்தப்படுகின்றன. மறைக்கப்பட்ட அடுக்குகள் அதிகரித்தால் செயல்திறன் குறைக்கப்படும்.

சில படைப்புகளில், கையால் எழுதப்பட்ட எழுத்துக்கள் உணர்தல் மாதிரியை (handwritten characters recognition model) ஆதரிக்க மறைக்கப்பட்ட மார்க்கோவ் மாதிரி (Hidden Markov Model) பயன்படுத்தப்பட்டது. இது அதிகபட்ச சாத்தியமான முடிவை வழங்குகிறது. பயிற்சி மாதிரியின் அருகிலுள்ள அண்டையர் வகைப்படுத்தி அணுகுமுறை (Nearest Neighbour classifier approach) பண்புக்கூறுகள் கணக்கிடப்பட்டு சேமிக்கப்படுகின்றன. இந்தப் பண்புக்கூறுகள் மாதிரிகள் சோதனைக்குப் பயன்படுத்தப்படுகின்றன. உள்ளீட்டு மாதிரிகள் சேமிக்கப்பட்ட மாதிரிகளுடன் ஒப்பிடப்படுகின்றன. சில படைப்புகளில், ஆதரவு திசையன் இயந்திரம் ஒரு இருமை வகைப்படுத்தியாகும், அங்கு பண்புக்கூறுகள் இரண்டு வகுப்புகளாகப் பிரிக்கப்படுகின்றன, அங்கு இரண்டு வகுப்புகளுக்கு இடையில் குறைந்தபட்ச விளிம்பின் மீத்தளத்துடன் பயன்படுத்தப்படுகிறது. இங்கே விளிம்பு, மீத்தளம் (hyper plane) மற்றும் நெருங்கிய தரவுப் புள்ளிகளுக்கு (closest data points) இடையிலான தூரத்தைக் குறிக்கிறது. விளைவு, தரவுப் புள்ளிகளை அடிப்படையாகக் கொண்டது, (அதாவது) விளிம்பில்; விளிம்பு அதிகபட்சமாக இருந்தால், மீத்தளம் துணை தளங்களாகப் பிரிக்கப்படும். நிகழ்வுகளை இரண்டு வகுப்புகளுக்கு மேல் வகைப்படுத்துவதில் சிக்கல் ஏற்பட்டால் (பல வகுப்பு சிக்கல்), பின்வரும் வழிமுறைகளைப் பயன்படுத்தவும், அனைவருக்கும் எதிராக ஒன்று மற்றும் ஒன்றுக்கு எதிராக.

தெளிவற்ற தர்க்க அணுகுமுறைகள் (Fuzzy logic approaches) மாதிரி தொடக்கநிலைகளை அடையாளம் காணவும் இரண்டு தொனி படத்தின் பெயரிடப்பட்டவையாகவும் பயன்படுத்தப்பட்டுள்ளன, பின்னர் இது ஒரு முன்மாதிரி எழுத்துக்களாக வகைப்படுத்தப்படும். இது புள்ளிவிவர வகைப்படுத்தியைப் பயன்படுத்தி எழுத்துக்களின் மேல் மற்றும் கீழ் ரகசிய வரம்புகளைக் காண்கிறது மேலும் அவை வகைப்படுத்தப்பட்ட விளைவாக சேமிக்கப்படுகின்றன. படிநிலை நரம்பியல் வலையமைப்பில் அணுகுமுறையில் (Hierarchical neural network approach) இரண்டு தொகுப்புகள் பயன்படுத்தப்படுகின்றன. முதல் குழு பல குழுக்களை வகைப்படுத்த பயிற்சி

அளிக்கப்படுகிறது. இரண்டாவது தொகுப்பில், ஒவ்வொரு குழுவும் மேலும் பல குழுக்களாகப் பிரிக்கப்படுகின்றன, அவற்றில் இருந்து, இறுதி முடிவு வகைப்படுத்தப்படுகிறது.

பல்வேறு சவால்களிலிருந்து பின்வரும் சவால்கள் அடையாளம் காணப்படுகின்றன, அவை இந்த பகுதியில் ஆராய்ச்சி பணிகளை மேற்கொள்ள ஆராய்ச்சியாளர்களுக்கு அதிக ஆர்வத்தை அளிக்கக்கூடும். அவை

- குறைந்த எண்ணிக்கையிலான தமிழ் எழுத்துக்கள் உணரப்பட்டுள்ளன (40 எழுத்துக்கள் அடையாளம் காணப்பட்டுள்ளன)
- பல்வேறு கையால் எழுதப்பட்ட ஆவணங்களிலிருந்து குறைந்தபட்ச சோதனை மாதிரிகள் பரிசீலிக்கப்பட்டுள்ளன
- அசாதாரண எழுத்து மற்றும் ஒத்த வடிவ எழுத்துக்களை அடையாளம் காண்பது கடினம்
- எழுத்துரு மாறுபாடு மற்றும் நெகிழ் எழுத்துக்கள் (sliding letters) - தீர்க்கப்படவில்லை
- உணர்வதில் குறைந்த துல்லிய விகிதம்
- பனை ஓலை, வரலாற்று ஆவணங்கள், தாய் ஆவணங்கள் மற்றும் கல் சிற்பங்கள் ஆகியவற்றில் கிடைக்கும் பழைய கையால் எழுதப்பட்ட எழுத்து தொகுப்பு இப்போது வரை தீர்க்கப்படவில்லை.

சேதமடைந்த பனை ஓலை மற்றும் சேதமடைந்த ஆவணங்களில் இருக்கும் சிதைந்த எழுத்துக்கள் இப்போது கருதப்படவில்லை.

முடிவு

தமிழ் கையால் எழுதப்பட்ட உணர்தலுக்கான ஆய்வில் நிறைய ஆராய்ச்சி பணிகள் உள்ளன. இருப்பினும், அனைத்து தமிழ் எழுத்துகளையும் நியாயமான துல்லியத்துடன் அடையாளம் காண நிலையான தீர்வு இல்லை. உணர்தல் செயல்பாட்டின் ஒவ்வொரு கட்டத்திலும் பல்வேறு முறைகள் பயன்படுத்தப்பட்டுள்ளன, அதேசமயம் ஒவ்வொரு அணுகுமுறையும் சில

எழுத்துக்குறி தொகுப்புகளுக்கு மட்டுமே தீர்வை வழங்குகிறது. உணர்தல் செயல்பாட்டின் போது இயல்பான மற்றும் அசாதாரண எழுத்து, சாய்ந்த எழுத்துக்கள், ஒத்த வடிவ எழுத்துக்கள், இணைந்த எழுத்துக்கள், வளைவுகள் மற்றும் பலவற்றை உணர்வதில் சவால்கள் இன்னும் உள்ளன. ஆஃப்லைன் தமிழ் எழுத்து உணர்தல் செயல்முறையின் ஒவ்வொரு கட்டத்தின் பல்வேறு அம்சங்களும் கணிக்கப்பட்டுள்ளன. ஆராய்ச்சியாளர்கள் குறைந்தபட்ச எழுத்துக்குறி தொகுப்பைப் பயன்படுத்தியுள்ளனர். வெவ்வேறு எழுத்து நடைகள் மற்றும் எழுத்துரு அளவு சிக்கல்களுக்கு பாதுகாப்பு வழங்கப்படவில்லை. பின்வரும் முக்கிய சவால்களை எதிர்காலத்தில் ஆராய்ச்சியாளர்கள் மேலும் ஆராயலாம்.

- தமிழ் எழுத்துக்களில் வளைவுகள்
- மிகப் பெரிய எழுத்துக்குறி தொகுப்பு
- சிக்கலான எழுத்து அமைப்பு
- சிக்கலான எழுத்து கட்டமைப்புகள் காரணமாக எழுதும் பாணிகளில் குறிப்பிடத்தக்க மாறுபாடு
- கோடுகள் மற்றும் துளைகளின் எண்ணிக்கை அதிகரித்தது
- கலப்புச் சொற்கள் (ஆங்கிலம் மற்றும் தமிழ்)
- தீவிர எழுத்துரு மாறுபாடு
- கோணங்கள், நிழல்கள் மற்றும் தனித்துவமான எழுத்துருக்களைப் பார்ப்பதில் உள்ள சிக்கல்கள்

தட்டச்சு செய்யப்பட்டு மற்றும் கையால் எழுதப்பட்டது

தட்டச்சு எழுதப்பட்டவற்றுடன் ஒப்பிடும்போது, கையெழுத்து உணர்தல் மிகவும் சவாலானது, ஏனெனில் எழுத்து அமைப்பு மற்றும் நோக்குநிலை அதை எழுதும் நபர்களின் பல்வேறு காரணிகளைப் பொறுத்தது. மேலும் கையெழுத்து அறிதல் புலம் ஆஃப்லைன் மற்றும்

ஆன்-லைன் உணர்தலாகப் பிரிக்கப்பட்டுள்ளது, அங்கு ஆஃப்லைன் உணர்தலில், கையெழுத்தின் படம் மட்டுமே கணினிக்கு கிடைக்கிறது, அதே நேரத்தில் ஆன்-லைன் வழக்கில் பேனா முனை போன்ற தற்காலிக தகவல்கள் நேரத்தின் செயல்பாடாக ஆயத்தொகுப்புகளும் கிடைக்கின்றன, இது முந்தையதை ஒப்பிடும்போது சற்று எளிதானது. ஆஃப்லைன் மற்றும் ஆன்-லைன் உணர்தலுக்கான பொதுவான தரவு கையகப்படுத்தும் சாதனங்கள் முறையே ஸ்கேனர்கள் மற்றும் டிஜிட்டல் மாத்திரைகள். மீண்டும் தற்காலிக தகவல் இல்லாததால், ஆஃப்லைன் கையெழுத்து அங்கீகாரம் ஆன்-லைனை விட கடினமாக கருதப்படுகிறது. மேலும், ஆஃப்லைன் வழக்கு என்பது மனிதனால் நிகழ்த்தப்படும் வழக்கமான எழுதும் பணிக்கு ஒத்ததாகும் என்பதும் தெளிவாகிறது.

தொழில்நுட்பங்கள் மற்றும் வழிமுறைகள்

எழுத்து உணரி ஒழுங்குமுறை உணர்தலின் வழிமுறைகளை விரிவாகப் பயன்படுத்துகிறது, இது அறியப்படாத மாதிரியை முன் வரையறுக்கப்பட்ட வகுப்பிற்கு ஒதுக்குகிறது. எழுத்து உணரிக்கான பல நுட்பங்களை வடிவ உணர்தலின் நான்கு பொது அணுகுமுறைகளில் ஆராயலாம்.

- வார்ப்புரு பொருத்தம் (Template Matching)
- புள்ளிவிவர நுட்பங்கள் (Statistical Techniques)
- கட்டமைப்பு நுட்பங்கள் (Structural Techniques)
- நரம்பியல் வலையமைப்புகள் (Neural Networks)

மேற்கண்ட அணுகுமுறைகள் ஒருவருக்கொருவர் சுதந்திரமாகவோ அல்லது முரண்படவோ இல்லை. எப்போதாவது, ஒரு அணுகுமுறையில் ஒரு எழுத்து உணரி நுட்பம் மற்ற அணுகுமுறைகளின் உறுப்பினராகவும் கருதப்படலாம்.

1) வார்ப்புரு பொருத்தம்

உணரப்பட வேண்டிய எழுத்து அல்லது வார்த்தைக்கு எதிராக சேமிக்கப்பட்ட முன்மாதிரிகளுடன் பொருந்தக்கூடிய எழுத்தின் எளிய வழிகளில் இதுவும் ஒன்றாகும். வெறுமனே, பொருந்தும் செயல்பாடு பண்புக்கூறு இடைவெளியில் இரண்டு திசையன்களுக்கு (பிக்சல்கள், வடிவங்கள், வளைவு போன்றவை) இடையிலான ஒற்றுமையின் அளவை தீர்மானிக்கிறது.

நேரடிப் பொருத்தம்

ஒரு சாம்பல்-நிலை அல்லது இருமை உள்ளீட்டு (gray-level or binary input character) எழுத்து நேரடியாக சேமிக்கப்பட்ட முன்மாதிரிகளின் நிலையான தொகுப்போடு ஒப்பிடப்படுகிறது. ஒரு வார்ப்புரு பொருத்தி (template matcher) பல தகவல் ஆதாரங்களை இணைக்க முடியும், இதில் போட்டி வலிமை மற்றும் வெவ்வேறு அளவீடுகளிலிருந்து அருகிலுள்ள அண்டை அளவீடுகள். நேரடிப் பொருந்தும் முறை உள்ளுணர்வு மற்றும் திடமான கணிதப் பின்னணியைக் கொண்டிருந்தாலும், இந்த முறையின் உணர்தல் விகிதம் சத்தத்திற்கு மிகவும் உணர்வுள்ளதாக (பாதிப்பு) இருக்கிறது.

சிரோமனி (Siromoney) குறியிடப்பட்ட எழுத்துச் சரம் அகராதியைப் பயன்படுத்தித் தமிழ் அச்சிடப்பட்ட எழுத்துக்களை உணரும் முறையை விவரித்தார். எழுத்து மேட்ரிக்ஸின் வரிசை மற்றும் நெடுவரிசை வாரியாக ஸ்கேனிங் மூலம் பிரித்தெடுக்கப்பட்ட சரம் பண்புக்கூறுகளை இந்த திட்டம் பயன்படுத்துகிறது. ஒவ்வொரு வரிசையிலும் உள்ள பண்புக்கூறுகள் உணரப்பட வேண்டிய ஸ்கிரிப்டின் சிக்கலைப் பொறுத்து பொருத்தமான முறையில் குறியாக்கம் செய்யப்படுகின்றன. கொடுக்கப்பட்ட உரை சின்னமாக குறியீடாக வழங்கப்படுகிறது மற்றும் ஒவ்வொரு சின்னத்திலிருந்து தகவல்களும் சரம் வடிவில் பிரித்தெடுக்கப்பட்டு அகராதியில் உள்ள சரங்களுடன் ஒப்பிடப்படுகின்றன. ஒரு ஒப்பந்தம் இருக்கும்போது, எழுத்துக்கள் ஒரு சிறப்பு ஒலிபெயர்ப்பைப் பின்பற்றி ரோமானிய எழுத்துக்களில் உணரப்பட்டு அச்சிடப்படுகின்றன.

தளர்வு பொருத்தம்

சின்னசுவாமி (Chinnuswamy) கையால் எழுதப்பட்ட தமிழ் எழுத்து உணர்தலுக்கான அணுகுமுறையை முன்மொழிந்தார். எழுத்துகள் தொடக்கநிலையானவைகள் எனப்படும் கூறுகள் போன்ற வரியால் ஆனதாகக் கருதப்படுகின்றன, சில தொடர்புடைய தடைகளை பூர்த்தி செய்கின்றன. கதாபாத்திரங்களின் கட்டமைப்பு அமைப்பை தொடக்கநிலைகளின் அடிப்படையில் விவரிக்க லேபிளிடப்பட்ட வரைபடங்கள் பயன்படுத்தப்படுகின்றன மற்றும் அவை திருப்தி அளிக்கும் தொடர்புடைய தடைகள். அங்கீகார செயல்முறை உள்ளீட்டு படத்தை உள்ளீட்டு தன்மையைக் குறிக்கும் பெயரிடப்பட்ட வரைபடமாக மாற்றுவதையும், அடிப்படை சின்னங்களின் தொகுப்பிற்காக சேமிக்கப்பட்ட பெயரிடப்பட்ட வரைபடங்களுடன் தொடர்பு குணகங்களை கணக்கிடுவதையும் கொண்டுள்ளது. இந்த வழிமுறை தொடர்பு குணகங்களை கணக்கிட இடவியல் பொருந்தும் நடைமுறையைப் பயன்படுத்துகிறது, பின்னர் தொடர்பு குணகத்தை அதிகரிக்கிறது.

சிதைக்கக்கூடிய வார்ப்புருக்கள் மற்றும் மீள் பொருத்தம் (Deformable Templates and Elastic Matching)

மீள் பொருத்துதல் நுட்பம் ஆஃப்லைன் மற்றும் ஆன்லைன் உணர்தல் அமைப்புகளுக்கு பயன்படுத்தப்படுகிறது. பிரசாந்த் மற்றும் பிறர் ஆன்லைன் தமிழ் எழுத்துக்களை உணர பிரபலமான மீள் பொருந்தும் வழிமுறையான டைனமிக் டைம் மடக்குதலைப் (Dynamic Time Wrapping) பயன்படுத்தினர். ஒரு நேரத் தொடரை அதன் நேர அச்சில் நீட்டிப்பதன் மூலமோ அல்லது சுருக்கியதன் மூலமோ நேர்கோட்டில் போர்த்த முடியுமானால், இது இரண்டு நேரத் தொடர்களுக்கு இடையில் உகந்த சீரமைப்பைக் கண்டறியும் ஒரு நுட்பமாகும். இரண்டு நேரத் தொடர்களுக்கிடையேயான இந்த மடக்குதல் அவற்றுக்கிடையேயான ஒற்றுமையைக் கண்டறிய பயன்படுத்தப்படலாம். சாகோ-சிபா பாண்ட் கட்டுப்பாட்டைப் (Sakoe-Chiba band constraint) பயன்படுத்துவதன் மூலம் இந்த முறை மிகவும் திறமையாகவும் வேகமாகவும் செய்யப்பட்டது.

வகைப்படுத்தலுக்குப் பயன்படுத்தப்படும் வெவ்வேறு பண்புக்கூறுகளி தொகுப்புகள்

• x-y பண்புக்கூறுகள் (எழுத்துக்களின் x மற்றும் y ஒருங்கிணைப்புகள்)

டேன்ஜென்ட் ஆங்கிள் (Tangent Angle (TA/டிஏ) பண்புக்கூறுடன் இணைந்து வடிவ சூழல் (Shape Context (SC/எஸ்சி) பண்புக்கூறு

• பொதுவான வடிவ சூழல் (Generalized Shape Context (GSC/ஜி.எஸ்.சி) பண்புக்கூறு

• x மற்றும் y மற்றும் வளைவு பண்புக்கூறுகளைப் பொறுத்து முதல் மற்றும் இரண்டாவது ஆக்கங்களை இயல்பாக்குகிறது.

மேலே குறிப்பிடப்பட்ட பண்புக்கூருகள் சோதனைக்கு பயன்படுத்தப்பட்ட நான்கு தொகுப்புகளை உருவாக்குகின்றன.

முதல் தொகுப்பு - x-y பண்புகூறுகள் மட்டும்

இரண்டாவது தொகுப்பு - டேன்ஜென்ட் ஆங்கிளானது வடிவ சூழல் உடன் சேர்க்கப்பட்டுள்ளது

மூன்றாவது செட் - பொதுவான வடிவ சூழல் மட்டும்

நான்காவது தொகுப்பு - 7 பண்புக்கூறுகள். முன் செயலாக்கப்பட்ட x-y பண்புக்கூறுகள், இயல்பாக்கப்பட்ட முதல் மற்றும் இரண்டாவது ஆக்கங்கள் மற்றும் வளைவு. (இடம்சார் பண்புக்கூறுகள் எல் 7)

டி.டி.டபிள்யூ (DTW) முடிவுகளை 4 வெவ்வேறு தொகுப்பு அம்சங்களுடன் ஒப்பிடுகையில், வடிவ சூழல் மற்றும் டேன்ஜென்ட் ஆங்கிள் பண்புகூறுகள் மெதுவாக காணப்படுகின்றன. துல்லியத்தின் அடிப்படையில் வடிவ சூழல் + டேன்ஜென்ட் ஆங்கிளை விட பொதுவான வடிவ சூழல் பண்புக்கூறுகள் சிறந்தது. இறுதியாக உள்ளூர் 7 பண்புக்கூறுகள் துல்லியம் மற்றும் வேகத்தின் அடிப்படையில் சிறந்த பண்புக்கூறுகளாக இருக்கின்றன.

நிரஞ்சன் மற்றும் பிறர் ஆன்லைன் தமிழ் எழுத்துக்களை உணர்வதில் துணைவெளி முறை (subspace method) மற்றும் டி.டி.டபிள்யூ (DTW) போன்ற இரண்டு வெவ்வேறு அணுகுமுறைகளின் சோதனைகளை ஒப்பிடுகின்றனர். துணைவெளி அடிப்படையிலான வகைப்பாடு (Subspace based

classification) அடிப்படையில் பண்புக்கூறு இடத்தின் (feature space) நேரியல் மாற்றமாகும். மாறுபாடு குறிப்பிடத்தக்கதாக இருக்கும் முதன்மை திசைகளைத் தேர்ந்தெடுப்பதன் மூலம், பண்புகூறு இடத்தை குறைந்த வரிசை இடத்தால் (order space) தோராயமாக மதிப்பிட முடியும். ஒவ்வொரு எழுத்துக்குறி வகுப்பும் துணைவெளியாக வடிவமைக்கப்பட்டுள்ளது. ஒரு முக்கிய முறை மற்றும் அதன் மாறுபாடுகள் காணப்படும்போது, இந்த வடிவங்களின் அனைத்து நேரியல் சேர்க்கைகளும் வகுப்பின் உறுப்பினர்களாகக் கருதப்படுகின்றன.

டி.டி.டபிள்யூ என்பது ஒரு மீள் பொருந்தக்கூடிய நுட்பமாகும், மேலும் இது வரிசைகளின் நேரியல் அல்லாத வரிசையமைப்பை (non-linear alignment of sequences) அனுமதிக்கிறது மற்றும் மிகவும் அதிநவீன ஒற்றுமை அளவைக் கணக்கிடுகிறது, இது முன்னேற்ற விகிதமானது நேரியல் அல்லாத முறையில் மாறுபடும் வடிவங்களை ஒப்பிடுவதற்கு மிகவும் பயனுள்ளதாக இருக்கும், இது யூசில்டியன் தூரம் (Eucildean distance) மற்றும் குறுக்கு தொடர்பு பயன்படுத்த முடியாதது (cross-correlation unusable) போன்ற ஒற்றுமை அளவை உருவாக்குகிறது.

இரண்டு முறைகளின் செயல்திறன் மூன்று வெவ்வேறு முறைகளுக்கு ஒப்பிடப்படுகிறது

1. எழுத்தாளர் சுதந்திர (Writer Independent WI),

2. எழுத்தாளர் சார்பு (Writer Dependent WD) மற்றும்

3. எழுத்தாளர் தகவமைப்பு (Writer Adaptive WA).

டி.டி.டபிள்யூ மற்றும் சப்ஸ்பேஸ் முறையின் செயல்திறனை ஒப்பிடுகையில், பிழைகள் குறித்த மூன்று வெவ்வேறு விளக்கங்கள் பின்வருமாறு.

குழு 1: இரண்டு முறைகளுக்கும் பொதுவான பிழைகள்

குழு 2: துணைவெளிக்கு குறிப்பிட்ட பிழைகள்

குழு 3: டி.டி.டபிள்யூ-க்கு குறிப்பிட்ட பிழைகள்

அட்டவணை 2: 3 வெவ்வேறு முறைகளின் செயல்திறன்

டி.டி.டபிள்யூ அடிப்படையிலான முறைகளின் செயல்திறன் ஓரளவு சிறப்பாக இருந்தாலும், வேகத்தைப் பொறுத்தவரை, துணைவெளி அடிப்படையிலான முறைகள் நன்றாக இருப்பதாகத் தெரிகிறது.

2) புள்ளிவிவர நுட்பங்கள் (Statistical Techniques)

புள்ளிவிவர முடிவுக் கோட்பாடு புள்ளிவிவர முடிவு செயல்பாடுகள் மற்றும் உகந்த அளவுகோல்களின் தொகுப்போடு தொடர்புடையது, இது குறிப்பிட்ட வகுப்பின் மாதிரியைக் கருத்தில் கொண்டு, கவனிக்கப்பட்ட வடிவத்தின் நிகழ்தகவை அதிகரிக்கிறது.

மறைக்கப்பட்ட மார்க்கோவ் மாதிரி (Hidden Markov Model (HMM))

கையால் எழுதப்பட்ட எழுத்து உணர்தல் சிக்கலுக்கு HMM மிகவும் பரவலாக மற்றும் வெற்றிகரமாக பயன்படுத்தப்படும் நுட்பமாகும். இது ஒரு ஒத்திசைவான மாதிரி என்பதால், எச்.எம்.எம் சத்தம் மற்றும் கையெழுத்தில் உள்ள மாறுபாடுகளை சமாளிக்க முடியும்.

பரத் மற்றும். பலர் (Bharath et. al.). தமிழ் எழுத்துக்களுக்கான தரவு உந்துதல் எச்.எம்.எம். (data-driven HMM) அடிப்படையிலான ஆன்லைன் கையால் எழுதப்பட்ட சொல் உணர்தல் முறையை முன்மொழிந்தது.

எழுதப்பட்ட எழுத்துக்களின் முன் செயலாக்கம் இரண்டு படிகளை உள்ளடக்கியது: சத்தம் நீக்குதல் மற்றும் இயல்பாக்குதல்.

முன் செயலாக்கத்திற்குப் பிறகு, பின்வரும் பண்புக்கூறுங்கள் போன்றவை சின்னங்களிலிருந்து பிரித்தெடுக்கப்படுகின்றன.

- இயல்பாக்கப்பட்ட Y. (Normalized Y)
- இயல்பாக்கப்பட்ட ஆக்கங்கள் (Normalized Derivatives)
- கோண பண்புக்கூகள் (Angle Features)
- பென்-அப் / பென்-டவுன் பிட் (Pen-up/Pen-down Bit)

இரண்டு சொல் மாதிரிகளை உருவாக்க HMM பயன்படுத்தப்படுகிறது: 1. சின்னம் மாடலிங் (Symbol modelling) மற்றும் 2. பென்-அப் ஸ்ட்ரோக் மாடலிங் (Pen-up Stroke modeling).

1. **குறியீட்டு மாடலிங்** (Symbol modeling) - ஒவ்வொரு சின்னத்தையும் மாடலிங் செய்வதற்கு, நிலையைத் தவிர்ப்பது இல்லாத எளிய இடமிருந்து வலமாக டோபாலஜி ஏற்றுக்கொள்ளப்பட்டது கைப்பற்றப்பட்டது. பின்னர் அது பாம்-வெல்ச் மறு மதிப்பீட்டு நடைமுறையைப் பயன்படுத்தி பயிற்சி அளிக்கப்படுகிறது. குறியீட்டின் வடிவ சிக்கலின் அடிப்படையில் ஒரு மாதிரிக்கு மாநிலங்களின் எண்ணிக்கை தீர்மானிக்கப்பட்டது.

2. **பென்-அப் ஸ்ட்ரோக் மாடலிங்** (Pen-up stroke modeling) - ஒரு குறியீட்டிற்குள் உள்ள பென்-அப் கோடுகள் குறியீட்டு மாதிரிகளைப் பயன்படுத்தி மறைமுகமாக வடிவமைக்கப்பட்டும்; அதேசமயம் சின்னங்களுக்கு இடையிலான பென்-அப் நிர்ணயிக்கப்படுகின்றது வெளிப்படையாக வடிவமைக்கப்படும். பொதுவான பென்-அப் ஸ்ட்ரோக் மாதிரிகள் கட்டப்பட்டவை; அவை எந்த ஜோடி சின்னங்களுக்கிடையில் பகிரப்பட்டும்; மேலும் அவை மாதிரிகள் என்ற வார்த்தையிலிருந்து பெறப்பட்ட இடைநிலைச் சின்னம் பென்-அப் கோடுகள் கொத்து (inter symbol pen-up strokes) மூலம் தீர்மானிக்கப்படுகின்றன. ஒவ்வொரு கொத்திலும்/கிளஸ்டரிலும் விழும் மாதிரிகள் ஒரு நிலைக்கு 2 காஸியர்களைக் கொண்ட (2 Gaussians per state) இரண்டு-நிலை இடமிருந்து வலமாக பென்-அப் எச்.எம்.எம் (two-state left-to-right pen-up HMM). அகராதியில் கொடுக்கப்பட்ட ஒரு சொல்லுக்கு, அதன் சொல் மாதிரியானது தொகுதி குறியீட்டு மாதிரிகளை ஒன்றிணைப்பதன் மூலமும் அவற்றுக்கு இடையில் செருகப்பட்ட பென்-அப் மாதிரிகளின் இணையான வலையமைப்பைக் கொண்டிருப்பதன் மூலமும் கட்டப்பட்டது. ஒரு அகராதி HMMகளின் ஒரு பிணையமாக குறிப்பிடப்படுகிறது, அங்கு தொடக்க முனையிலிருந்து இறுதி முனை வரை பிணையத்தின் ஒவ்வொரு பாதையும் ஒரு வார்த்தையுடன் ஒத்திருக்கும். சிறந்த விட்டெர்பி டிகோடிங் (standard Viterbi decoding) மூலம் சிறந்த பாதை மதிப்பீடு தீர்மானிக்கப்படும்.

சொல் உணர்தல் முறையின் மதிப்பீடு செயல்திறனை மதிப்பிடுவதற்கு 1K, 2K, 5K, 10K மற்றும் 20K போன்ற வெவ்வேறு அகராதி அளவுகளில் மேற்கொள்ளப்படும்.

அட்டவணை 3: வெவ்வேறு லெக்சிகன் அளவுகளுக்கான செயல்திறன் துல்லியம்

ஜெகதீஷ் கண்ணன் மற்றும் பிரபாகர் (Jagadeesh Kannan & Prabhakar) ஆஃப்-லைன் கர்சீவ் கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களை உணர்வதற்காக தனித்துவமான மறைக்கப்பட்ட மார்க்கோவ் மாதிரிகளைப் பயன்படுத்துகின்றனர். ஒவ்வொரு எழுத்துக்கும் இரண்டு எச்.எம்.எம் கள் உருவாக்கப்படுகின்றன, ஒன்று கிடைமட்ட தகவல்களை மாடலிங் செய்வதற்கும் மற்றொன்று முறையே செங்குத்து தகவல்களை மாடலிங் செய்வதற்கும்.

HMMக்கான அதிகபட்ச நிகழ்தகவு அளவுரு மதிப்பீடு பல கண்காணிப்பு வரிசைகளுடன், பாம்-வெல்ச் வழிமுறை என்ற செயல்பாட்டு செயல்முறையால் பெறப்படுகிறது.

தனித்துவமான மறைக்கப்பட்ட மார்கோவ் எழுத்துக்கள் மாதிரி நிலையான நடைமுறைகளைப் பயன்படுத்தி பயிற்சியளிக்கப்படுகிறது, மேலும் அனைத்து எழுத்து HMMகளுக்கும் மாநிலத்தின் எண்கள் சரி செய்யப்படுகின்றன மற்றும் தவிர்க்கும் நிலைகள் அனுமதிக்கப்படவில்லை. இரண்டு பதிவு

தனித்துவமான மறைக்கப்பட்ட மார்க்கோவ் எழுத்துக்கள் மாதிரி நிலையான நடைமுறைகளைப் பயன்படுத்தி பயிற்சியளிக்கப்படுகிறது, மேலும் அனைத்து எழுத்துக்குறி எச்.எம்.எம்.-களுக்கும் நிலையின் எண்கள் சரி செய்யப்படுகின்றன, மேலும் தவிர்க்கும் நிலைகள் அனுமதிக்கப்படாது. ஒவ்வொரு வேட்பாளருக்கும் இரண்டு பதிவு நிகழ்தகவுகள் கிடைமட்ட திசை HMM-ஐப் பயன்படுத்தி கணக்கிடப்படுகின்றன. பின்னர், இறுதி 3-சிறந்த எழுத்து அங்கீகாரத்தைப் பெற பதிவு நிகழ்தகவுகள் ஒன்றாக சேர்க்கப்படுகின்றன

முன்மொழியப்பட்ட அமைப்பு வெவ்வேறு நபர்களின் பல கையால் எழுதப்பட்ட ஆவணங்கள், ஒலிச்சுவாடி [37] மாதிரிகள், இயந்திரம் அச்சிடப்பட்ட, ஸ்கேன் செய்யப்பட்ட ஆவணங்களுடன் சோதிக்கப்பட்டது, அவை பின்வரும் முடிவுகளைத் தருகின்றன.

அட்டவணை 4: செயல்திறன் அட்டவணை

நிகழ்தகவு இல்லாத மாதிரிகள்

ஆதரவு திசையன் இயந்திரங்கள் (Support Vector Machines) என்பது தரவுகளிலிருந்து கற்றல், வகைப்பாடு மற்றும் பின்னடைவு விதிகளுக்கான பயிற்சி வழிமுறையாகும். செயற்கை அறிவில் திறமையான (Aritificial Intelligence) கற்றல் முறைகளில் இதுவும் ஒன்றாகும். எஸ்.வி.எம் என்பது பின்வரும் கூறுகளை உள்ளடக்கிய ஒரு நிகழ்தகவு இல்லாத மாதிரி (Non-probabilistic model):

• ஒழுங்குபடுத்தப்பட்ட நேரியல் கற்றல் மாதிரிகள் (வகைப்பாடு மற்றும் பின்னடைவு போன்றவை),

• கோட்பாட்டு எல்லைகள்,

•குவிந்த இருமை மற்றும் அதனுடன் தொடர்புடைய இரட்டை-கர்னல் பிரதிநிதித்துவம், மற்றும்

• இரட்டை கர்னல் பிரதிநிதித்துவத்தின் இடைவெளி.

இது எஸ்.வி.எம் மற்றும் தொடர்புடைய கர்னல் அடிப்படையிலான கற்றல் முறைகளை சிறப்பு மற்றும் சுவாரஸ்யமாக்குகிறது. எஸ்.வி.எம்-களும் நடைமுறையில் வெற்றிகரமாக பயன்படுத்தப்படுகின்றன, குறிப்பாக வகைப்பாடு சிக்கல்களுக்கு. எஸ்.வி.எம்.-களால் வெற்றிகரமாக தீர்க்கப்பட்ட பல சிக்கல்களும் நிலையான புள்ளிவிவர முறைகளால் வெற்றிகரமாக தீர்க்கப்பட்டிருக்கலாம்.

சீதலட்சுமி மற்றும் பிறர்-இல் (Shanthi et. al.) அச்சிடப்பட்ட தமிழ் உரை ஆவணங்களை மென்பொருளாக மொழிபெயர்க்கப்பட்ட யூனிகோட் தமிழ் உரையாக மாற்றும் செயல்முறையைக் குறிக்கும் OCR ஐ உருவாக்குகிறது.

ஒவ்வொரு எழுத்துக்குறி கிளிஃபுக்கும் ஒத்ததாக பிரித்தெடுக்கப்பட்ட பின்வருவன போன்ற பண்புக்கூறுகளைப் பயன்படுத்தி வகைப்பாடு செய்யப்படுகிறது.

• எழுத்தின் உயரம் மற்றும் அகலம்

• கிடைமட்ட கோடுகள்-குறுகிய மற்றும் நீண்ட

• செங்குத்து கோடுகள்-குறுகிய மற்றும் நீண்ட

• வட்டங்களின் எண்ணிக்கை

•கிடைமட்டமாக சார்ந்த மற்றும் செங்குத்தாக சார்ந்த வளைவுகளின் எண்ணிக்கை

Etc. படத்தின் சென்ட்ராய்டு போன்றவை,

வகைப்படுத்தல் ஆதரவு திசையன் இயந்திரங்களால் செய்யப்படுகிறது. பயன்படுத்தப்படும் கர்னல் செயல்பாடு RBF (ரேடியல் பேஸிஸ் செயல்பாடு) ஆகும், இது மிகவும் பிரபலமான தேர்வாகும். எஸ்.வி.எம் ஒரு கற்றல் தொகுதி மற்றும் ஒரு வகைப்பாடு தொகுதி ஆகியவற்றைக் கொண்டுள்ளது. பயிற்சி மாதிரி உள்ளீட்டு கோப்பு, இலக்கு கோப்பை எடுத்து பிணையத்திற்கு (network) பயிற்சி அளிக்கிறது. வகைப்பாடு மாதிரியில், வகுப்பு 1, 2, 3, …… 247 போன்ற பல்வேறு வகுப்பு லேபிள்கள் கொடுக்கப்பட்டுள்ளன. இவ்வாறு எஸ்.வி.எம் வகுப்புகளின் சரியான லேபிள்களைக் கற்றுக் கொண்டு உற்பத்தி செய்கிறது. வகைப்படுத்தலுக்குப் பிறகு எழுத்துக்கள் உணரப்பட்டு ஒரு மேப்பிங் அட்டவணை உருவாக்கப்படுகிறது, அதில் தொடர்புடைய எழுத்துக்களுக்கான யூனிகோட்கள் மேப் செய்யப்படுகின்றன. பல்வேறு செயல்பாட்டுத் தொகுதிகள் மூலம் ஸ்கேன் செய்யப்பட்ட படம் இறுதியாக மேப்பிங் அட்டவணையில் இருந்து அங்கீகரிக்கப்பட்ட விவரங்களுடன் ஒப்பிடப்படுகிறது, அதிலிருந்து தொடர்புடைய யூனிகோட்கள் அணுகப்பட்டு யூனிகோட் எழுத்துருக்களைப் பயன்படுத்தி அச்சிடப்படுகின்றன, இதனால் OCR அடையப்படுகிறது.

சாந்தி மற்றும் பிறரில் ஆதரவு திசையன் இயந்திரம் (எஸ்.வி.எம்) அடிப்படையில் ஆஃப்லைன் கட்டுப்படுத்தப்படாத கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களுக்கான அங்கீகார முறையை விவரித்தார். எஸ்.வி.எம். என்பது ஒரு நாவல் புள்ளிவிவர கற்றல் நுட்பத்தை அடிப்படையாகக் கொண்ட ஒரு வகை வகைப்படுத்தியாகும். கையால் எழுதப்பட்ட

கதாபாத்திரங்களில் பெரும் மாறுபாட்டில் உள்ள சிரமம் காரணமாக, இந்த அமைப்பு 106 எழுத்துக்களுடன் பயிற்சியளிக்கப்பட்டு தேர்ந்தெடுக்கப்பட்ட 34 தமிழ் எழுத்துக்களுக்கு சோதிக்கப்படுகிறது.

மாதிரி தரவு தொகுப்பு கிட்டத்தட்ட அனைத்து எழுத்துக்களையும் குறிக்கும் வகையில் தரவு தொகுப்பு தேர்ந்தெடுக்கப்பட்டது. A4 அளவிலான ஆவணங்களில் வெவ்வேறு எழுத்தாளர்களிடமிருந்து தரவு மாதிரிகள் சேகரிக்கப்படுகின்றன. அவை 300 டிபிஐ தீர்மானத்தில் ஒரு தட்டையான படுக்கை ஸ்கேனரைப் பயன்படுத்தி ஸ்கேன் செய்யப்பட்டு சாம்பல் அளவிலான படங்களாக சேமிக்கப்படுகின்றன. படத்தின் தரத்தை மேம்படுத்த டிஜிட்டல் மயமாக்கப்பட்ட படத்தில் பல்வேறு முன் செயலாக்க நடவடிக்கைகள் முன்னரே வடிவமைக்கப்பட்டுள்ளன.

சீரற்ற அளவிலான முன் செயலாக்கப்பட்ட படம் சீரான அளவிலான படத்திற்கு இயல்பாக்கப்படுகிறது. படத்தின் வெவ்வேறு மண்டலங்களுக்குப் பிக்சல் அடர்த்தி கணக்கிடப்படுகிறது மற்றும் இந்த மதிப்புகள் ஒரு எழுத்தின் பண்புக்கூறுகளாகப் பயன்படுத்தப்படுகின்றன. ஆதரவு திசையன் இயந்திரத்தை பயிற்றுவிக்கவும் சோதிக்கவும் இந்தப் பண்புக்கூறுகள் பயன்படுத்தப்படுகின்றன.

சிவ்சுப்பிரமணி மற்றும் பிறர் (Shivsubramani et. al) [14] எழுத்துக்களுக்கு இடையிலான உறவை ஆராயும் அச்சிடப்பட்ட தமிழ் எழுத்துக்களை உணர்வதற்கான ஒரு முறையை முன்வைத்தார். மல்டி-கிளாஸ் ஹைரார்ச்சிகல் சப்போர்ட் வெக்டர் மெஷின் (Multi-class Hierarchical Support Vector Machine) என்பது உணர்தலுக்கான மல்டி-கிளாஸ் சப்போர்ட் வெக்டர் மெஷினின் மாறுபாடு ஆகும்.

இங்கே தனிப்பட்ட எழுத்துக்கள் ஆராயப்படுகின்றன மற்றும் அவற்றின் வடிவங்களின் அடிப்படையில் நிறைய இடை-வகுப்பு சார்புகள் அடையாளம் காணப்படும், இது எழுத்துக்களை உணர்வதற்கான படிநிலைகளாக எழுத்துக்களை ஒழுங்கமைக்க உதவுகிறது.

ஒத்த எழுத்துக்களின் பண்புக்கூறுமதிப்புகள் மிகக் குறைந்த வேறுபாட்டைக் கொண்டுள்ளன. ஒற்றுமையை வெளிப்படுத்தும் எழுத்துக்கள் வகைப்படுத்தலுக்கான படிநிலைகளாக ஒழுங்கமைக்கப்பட்டன. ஒரு குறிப்பிட்ட துணைப்பிரிவின் பண்பை வெளிப்படுத்த ஒரு எழுத்து பண்புக்கூறு காணப்படுகிறது, மேலும் இந்த வரிசைக்கு சொந்தமில்லாத எழுத்துக்கள் வகைப்படுத்தலுக்கு கருதப்பட வேண்டியதில்லை. இதன்மூலம் 126 வகுப்பு வகைப்பாடு சிக்கல் 10 வகுப்பு அல்லது 8 வகுப்பு சிக்கலாக பிரிக்கப்படுகிறது. இது மல்டி லேயர் பெர்செப்டிரான் (Multi-Layer Perceptron (MLP/எம்.எல்.பி), கே-அருகிலுள்ள அண்டை (K-Nearest neighbor (KNN/கே.என்.என்) மற்றும் முடிவு கிளைகள் (Decision Trees(DT/.டி.டி) ஆகியவற்றை விடத் திறமையானது என்று கண்டறியப்பட்டுள்ளது.

ஜெகதீஷ் மற்றும் பலர் (Jagadeesh et. al) [21] தமிழ் எழுத்துக்களை உணர்வதற்காக இரண்டு வழிமுறைகளை இணைக்கும் ஒரு அமைப்பை வழங்கினார். துல்லியத்தை அதிகரிக்க இரண்டு வழிமுறைகளின் நன்மை மற்றும் செயல்திறன் இணைக்கப்படுகின்றன.

முன் செயலாக்க நிலைக்குப் பிறகு, HMM மற்றும் SVM இன் இணைவைப் பயன்படுத்தி எழுத்துக்கள் உணரப்படுகின்றன.

இறுதியாக இரண்டு வழிமுறைகளின் வெளியீட்டைக் கொண்டு ஒரு ரேடியல் அடிப்படை செயல்பாட்டு நரம்பியல் நெட்வொர்க் (Radial Basis Functional Neural network) பயிற்சி பெற்றது. எச்.எம்.எம் அல்லது எஸ்.வி.எம் தவறான எழுத்துக்களைக் கொடுத்தால், நரம்பியல் நெட்வொர்க், வழிமுறைகள் மற்றும் உண்மையான எழுத்து ஆகிய இரண்டின் எடை வயதுடன் பயிற்சியளிக்கப்படுகிறது.

இந்த இணைவுக்குப் (fusion) பின்னால் உள்ள யோசனை

1. ஒரு வழிமுறை எழுத்தை அடையாளம் காணத் தவறினால், மற்றொரு வழிமுறை எழுத்தை அடையாளம் காண ஆதரிக்கக்கூடும்.

2. ஒரு வழிமுறை தவறான எழுத்தைக் கொடுத்தால், இன்னொன்று சரியான ஒன்றைக் கொடுக்கலாம்.

3. இரண்டு வழிமுறைகளாலும் ஒரே தவறான அடையாளத்திற்கான வாய்ப்பு குறைவாக உள்ளது.

4. ஒரு வழிமுறை தவறான முடிவைக் கொடுத்தால், சரியான முடிவைத் தேர்ந்தெடுக்கும் முடிவு நரம்பியல் பிணையத்தால் செய்யப்படுகிறது.

இரண்டு வழிமுறைகளின் தவறான உணர்தலுக்காக இந்த செயல்முறை செய்யப்படுகிறது. இரண்டு வழிமுறைகளும் ஒரே எழுத்தைக் கொடுக்காதபோது, உண்மையான எழுத்தை மீட்டெடுக்க பயிற்சி பெற்ற RBFNN பயன்படுத்தப்படுகிறது. அமைப்பின் செயல்திறனை சோதிக்க அதிகபட்சம் 250 திருக்குரள்கள் தேர்ந்தெடுக்கப்பட்டன.

தெளிவற்ற தொகுப்பு பகுத்தறிவு (Fuzzy Set Reasoning)

சுரேஷ் மற்றும் பிறர் (Suresh et. al.) [24] கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களில் தெளிவற்ற கருத்தை பயன்படுத்துவதற்கான அணுகுமுறையை விவரிக்கிறது, அவற்றை முன்மாதிரி எழுத்துக்களில் ஒன்றாக வகைப்படுத்தலாம், இது சட்டத்திலிருந்து தூரம் மற்றும் பொருத்தமான உறுப்பினர் செயல்பாடு (distance from the frame and a suitable membership function) என்று அழைக்கப்படுகிறது. அறியப்படாத மற்றும் முன்மாதிரி எழுத்துக்கள் முன் செயலாக்கப்பட்டு உணர்தலுக்காக கருதப்படுகின்றன. தெளிவற்ற தொகுப்பின் கோட்பாடு தவறான வரையறுக்கப்பட்ட அமைப்புகளின் நடத்தை விவரிக்க தோராயமான ஆனால் பயனுள்ள வழிமுறையை வழங்குகிறது. கையால் எழுதப்பட்ட எழுத்துக்கள் போன்ற மனித தோற்றத்தின் வடிவங்கள் ஓரளவிற்கு இயற்கையில் தெளிவில்லாமல் காணப்படுகின்றன. தெளிவற்ற கருத்தியல் அணுகுமுறையை திறம்பட பயன்படுத்த இது விவரிக்கப்பட்டுள்ளது. அல்காரிதம் எண்களுக்கு 250 மாதிரிகள் மற்றும் தேர்ந்தெடுக்கப்பட்ட ஏழு தமிழ் எழுத்துக்களுக்கு சோதிக்கப்படுகிறது மற்றும் பெறப்பட்ட வெற்றி விகிதம் 76% முதல் 94% வரை மாறுபடும்.

3) கட்டமைப்பு நுட்பங்கள்

ஜெகதீஷ் மற்றும் பலர் (Jagadeesh et. al) ஆஃப்லைன் தமிழ் எழுத்துக்களை உணர்தலுக்காக ஆக்டல் வரைபட (Octal graph conversion) மாற்றத்தைப் பயன்படுத்துகின்றனர். முன்மொழியப்பட்ட அணுகுமுறை உணர்வதற்கான ஒரு தீர்வை மேற்கொள்கிறது, இது கொடுக்கப்பட்ட எழுத்தின் ஒவ்வொரு பிக்சலையும் ஒரு வரைபடத்தின் முனையாகக் குறிப்பதன் மூலம் எழுதப்பட்ட எழுத்தை ஆக்டல் வரைபடமாக மாற்றுகிறது. ஒவ்வொரு கணுக்கும் எட்டு புலங்கள் உள்ளன, எனவே ஆக்டல் வரைபடம் (octal graph) என அழைக்கப்படுகிறது. எழுதும் பாணியிலிருந்து சுயாதீனமாக ஒரு கடிதத்தின் அடிப்படை வடிவத்தை குறிக்க வரைபடம் முயற்சிக்கிறது. வரைபடங்களின் எடையைப் பயன்படுத்தி, முன் வரையறுக்கப்பட்ட எழுத்துகளுடன் பொருந்தக்கூடிய பொருத்தமான அம்சத்தின் மூலம், எழுதப்பட்ட எழுத்துக்கள் அங்கீகரிக்கப்படுகின்றன.

ஒரு சாதாரண வரைபடத்தைப் போலல்லாமல் ஒரு ஆக்டல் வரைபடம் எட்டு சுட்டிகள் மற்றும் தரவு புலத்துடன் ஒரு முனை உள்ளது. அண்டை பிக்சல்களின் அடிப்படையில், ஆக்டர் முனையின் பல்வேறு புலங்களுக்கு சுட்டிக்காட்டி மதிப்புகள் ஒதுக்கப்படுகின்றன. இந்த ஆக்டல் முனைகள் வாசல் மதிப்பின் அடிப்படையில் மற்ற முனைகளுடன் இணைக்கப்பட்டுள்ளன. இந்த எண்ணிக்கை ஒரு தமிழ் உயிரெழுத்தின் ஆக்டல் நோட் (octal node) பிரதிநிதித்துவத்தைக் காட்டுகிறது.

4) நரம்பியல் நெட்வொர்க்குகள்

ஒரு நரம்பியல் நெட்வொர்க் ஒரு கம்ப்யூட்டிங் கட்டமைப்பாக வரையறுக்கப்படுகிறது, இது தகவமைப்பு 'நரம்பியல்' செயலிகளின் பாரிய இணையான இணைப்புகளைக் கொண்டுள்ளது. அதன் தகவமைப்பு தன்மை காரணமாக, இது தரவுகளில் ஏற்படும் மாற்றங்களுக்கு ஏற்ப மற்றும் உள்ளீட்டு சமிக்ஞையின் பண்புகளை அறிய முடியும். ஒரு நரம்பியல் பிணையத்தில் பல முனைகள்

உள்ளன. ஒரு முனையிலிருந்து வெளியீடு நெட்வொர்க்கில் இன்னொருவருக்கு வழங்கப்படுகிறது மற்றும் இறுதி முடிவு அனைத்து முனைகளின் சிக்கலான தொடர்புகளையும் சார்ந்துள்ளது. வேறுபட்ட அடிப்படைக் கொள்கைகள் இருந்தபோதிலும், பெரும்பாலான நரம்பியல் நெட்வொர்க் கட்டமைப்புகள் புள்ளிவிவர அமைப்பொழுங்கு உணர்தல் முறைகளுக்கு (statistical pattern recognition methods) சமமானவை என்பதைக் காட்டலாம் [38].

அபர்ணா மற்றும் பிறர் (Aparna et. al) [19] தமிழ் செய்தித்தாள் ஒரு OCR முறையை முன்வைக்கிறனர். பிரிவின் முக்கிய சிக்கல்களைத் தீர்க்க மற்றும் எழுத்து உணர்தலுக்காக ANN பயன்படுத்தப்படுகிறது.

எழுத்து உணர்தலுக்காக ஒரு ரேடியல் அடிப்படை செயல்பாடு நரம்பியல் பிணையத்தைப் பயிற்சி செய்கிறது. உயிரெழுத்துகள், மெய், சிறப்பு எழுத்து அல்லது கிரந்தா எழுத்துக்கள் மற்றும் 0 முதல் 9 வரையிலான ஆங்கில எண்கள் மற்றும் மொத்தம் 157 எழுத்துக்களை உருவாக்கும் நிறுத்தற்குறிகள் உள்ளிட்ட அனைத்து எழுத்துக்களும் நெட்வொர்க்கைப் பயிற்றுவிப்பதற்காக எடுக்கப்படுகின்றன. எழுத்துக்கள் 52 x 52 சாளரத்தின் மையத்தில் வைக்கப்பட்டுள்ளன, மேலும் RBF நரம்பியல் நெட்வொர்க்கால் பயிற்சியளிக்கப்பட வேண்டிய உள்ளீட்டு அமைப்பொழுங்குகள் 40 காபார் வடிப்பான்களில் (Gabor filters) ஒவ்வொன்றிலும் நான்கு திசைகளில் ஒவ்வொன்றிலும் 10 உடன் எழுத்து உடன் புள்ளி தயாரிப்பை எடுத்துக்கொள்வதன் மூலம் பெறப்படுகின்றன. RBF நரம்பியல் வலைப்பின்னல் ஒவ்வொரு வெளியீட்டையும் ஒரு எழுத்துக்களுடன் தொடர்புடைய 157 வெளியீடுகளைக் கொண்டுள்ளது.

இந்த பயிற்சி பெற்ற நரம்பியல் நெட்வொர்க் பொதுவாக 85 முதல் 90 சதவிகிதம் துல்லியத்தை அளிக்கும் பிரிக்கப்பட்ட எழுத்துக்களை அங்கீகரிப்பதற்கு பயன்படுத்தப்படுகிறது.

சுதா மற்றும் பிறர் (Sudha et. al) [26] ஒரு மறைக்கப்பட்ட அடுக்குடன் பல அடுக்கு உணர்வைப் பயன்படுத்தி கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களை உணர்வதற்கான

அணுகுமுறையை விவரிக்கிறது. கையால் எழுதப்பட்ட எழுத்திலிருந்து பிரித்தெடுக்கப்பட்ட அம்சம் ஃபோரியர் விளக்கங்கள். கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களை உணர்வதில் பின் பரப்புதல் வலையமைப்பின் உயர் செயல்திறனை அடைய மறைக்கப்பட்ட அடுக்கு முனைகளின் (hidden layer nodes) எண்ணிக்கையை தீர்மானிக்க ஒரு பகுப்பாய்வு மேற்கொள்ளப்பட்டது.

வெவ்வேறு வயதினரின் ஆண் மற்றும் பெண் பங்கேற்பாளர்களால் வழங்கப்பட்ட பல்வேறு வகையான கையெழுத்துக்களைப் பயன்படுத்தி இந்த அமைப்பு பயிற்சி பெற்றது. சோதனை முடிவுகள், பின் பரப்புதல் நெட்வொர்க்குடன் இணைந்து ஃபோரியர் விளக்கங்கள் கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களுக்கு 97% நல்ல அங்கீகார துல்லியத்தை வழங்குகின்றன.

செயல்திறன் பகுப்பாய்வு (Performance Analysis)

தமிழ் எழுத்து உணர்தல் துறையில் நடந்துகொண்டிருக்கும் ஆராய்ச்சியில் பயன்படுத்தப்பட்ட முக்கிய அணுகுமுறைகளின் கண்ணோட்டத்தை இதுவரை நாம் பார்த்தோம். மேலே விவாதிக்கப்பட்ட ஒவ்வொரு முறைகளும் அவற்றின் சொந்த மேன்மைகள் மற்றும் குறைபாடுகளைக் கொண்டிருந்தாலும், உணர்தல் துல்லியம் விகிதங்கள் 85%க்கு மேல் இருப்பதாகக் கூறப்படுகிறது. இருப்பினும், ஒவ்வொரு முறைகளின் வெற்றியை ஒப்பிடுவது மிகவும் கடினம், குறிப்பாக உணர்தல் விகிதங்களின் அடிப்படையில், வெவ்வேறு தரவுத்தளங்கள், கட்டுப்பாடுகள் மற்றும் மாதிரி இடங்கள் போன்ற பல்வேறு காரணிகளால்.

உயர்தர தாளில் அல்லது டேப்லெட்டில் தனித்தனியாகவும் சுத்தமாகவும் இருக்கும் நூல்களில் தற்போதைய ஆராய்ச்சி வெற்றிகரமாக உள்ளது, அதன் உணர்தல் விகிதங்கள் 85%-க்கும் அதிகமாக உள்ளன, மேலும் ஒரு சில ஆராய்ச்சி கர்சீவ் ஆஃப்லைன் நூல்களில் இருந்தது. குறைந்த தரம் வாய்ந்த காகிதத்தில் மோசமாக இருக்கும் கையால் எழுதப்பட்ட நூல்களுக்கு, இன்னும் தீவிர ஆராய்ச்சி தேவை.

பெரும்பாலான ஆஃப்லைன் நூல்கள் HMM-கள் மற்றும் நரம்பியல் நெட்வொர்க்குகளைப் பயன்படுத்தி உணரப்பட்டுள்ளன. வெவ்வேறு நுட்பங்களை இணைப்பதும் நல்லது, ஏனென்றால் ஒவ்வொரு நுட்பத்திலும் சிறந்தது உணர்தலுக்காகப் பயன்படுத்தப்படுகிறது, இது துல்லிய விகிதங்களை மேம்படுத்தக்கூடும்.

4. முடிவு

ஆன்லைன் மற்றும் ஆஃப்லைன் கையால் எழுதப்பட்ட மற்றும் தட்டச்சு செய்யப்பட்ட எழுத்துக்களுக்கான தமிழ் எழுத்துக்களுக்கான ஒளிவழி எழுட்துணரி (ஆப்டிகல் கேரக்டர் ரெக்னிகேஷன்) ஒழுங்குமுறைகளில் (சிஸ்டங்களில்) நடந்து வரும் ஆராய்ச்சியின் கண்ணோட்டத்தை அளிப்பதே நமது இலக்கு. பல்வேறு முறைகளின் செயல்திறன் பகுப்பாய்வு செய்யப்பட்டுள்ளது, இது ஒவ்வொரு முறையின் திறமையான காரணியின் படத்தை அளிக்கிறது.

தொகுப்புரை

தமிழுக்கு கூகுள் எழுத்துணரி தொழில் நுட்பம் பற்றி விரிவாகக் கூறப்பட்டது. கூகுள் எழுத்துணரி டுடோரியல் பற்றி ஒரு விளக்கம் தரப்பட்டது. அதைத் தொடர்ந்து தமிழ் கையெத்துப் படிவத்திற்கான எழுத்துணரித் தொழிநுட்பம் பற்றி விரிவாகக் கூறப்பட்டது. தமிழ் கையெத்துப் படிவத்திற்கான எழுத்துணரி உருவாக்கத்தில் ஆய்வாளர்களின் பங்களிப்பு பற்றிக் கூறப்பட்டது.

நோக்கீடுகள்

Antony Robert Raj, S.Abirami. A Survey on Tamil Handwritten Character Recognition using OCR Techniques. பதிவிறக்கம் 10.11.2020.

Aparna K. H., Sumanth Jaganathan, P. Krishnan, V. S Chakravathy. “Document Image Analysis with specific application to Tamil Newsprint”.

Banumathi P and Nasira G.M. 2011 “Handwritten Tamil Character Recognition using Artificial neural networks”, International Conference on Process Automation, Control and Computing (PACC), page(s): 1 – 5, 2011.

Bharath A, Sriganesh Madhvanath, 2007. “Hidden Markov Models for online handwritten Tamil word recognition”.

Bhattacharya U, Ghosh S.K and Parui S.K. 2007 “A Two Stage Recognition Scheme for Handwritten Tamil Characters”, Ninth International Conference on Document Analysis and Recognition, Vol: 1 page(s): 511 – 515, 2007.

Chinnuswamy. P and S G Krishnamoorthy 1980. “Recognition of hand printed Tamil characters”, Pattern Recognition, 12:141

Jagadeesh Kannan R., R. Prabhakar, 2008. “Off-Line cursive handwritten Tamil character recognition”.

R. Jagadeesh Kannan, R. Prabhakar, 2008. “An improved Handwritten Tamil character recognition system using Octal Graph”.

Jagadeesh Kannan R., R. Prabhakar, 2009. “A Comparative study of Optical Character Recognition for Tamil script”.

Niranjan Joshi, G. Sita and A. G. Ramakrishnan. “Compariso of Elastic matching algorithms for online Tamil handwritten character recognition”.

Niranjan joshi, G. Sita , A. G. Ramakrishnan. “Tamil handwriting recognition using subspace and DTW based classifiers”.

Shanthi N and Duraiswami K. 2010 “A Novel SVM-based Handwritten Tamil character recognition system”, springer, Pattern Analysis & Applications,Vol-13, No. 2, 173-180,2010.

Shivsubramani K, Loganathan R, Srinivasan C J, Ajay V, Soman K P, 2007. “Multiclass Hierarchical SVM for recognition of printed Tamil characters”.

Stuti Asthana, Farha Haneef and Rakesh K Bhujade, “Handwritten Multiscript Numeral Recognition using Artificial Neural Networks”, Int. J. of Soft Computing and Engineering ISSN: 2231-2307, Volume-1, Issue-1, March 2011.

Seethalakshmi R, Sreeranjani T R, Balachandar T. “Optical Character Recognition for printed Tamil text using Unicode”.

Suresh Kumar C and Ravichandran T. 2010 “Handwritten Tamil Character Recognition using RCS algorithms”, Int. J. of Computer Applications, (0975 – 8887) Volume 8– No.8, October 2010.

Sutha, J Ramaraj, N Sethu. “Neural Network based Offline Tamil handwritten character recognition system”.

Siromoney et al., 1978. “Computer recognition of printed Tamil character”, Pattern Recognition 10:243-247.

வினாவங்கி

1.பொருத்தமான விடையைத் தேர்ந்தெடுத்தல் (15 வினாக்கள்)

1) ஒளிவழி எழுத்துணரி அமைப்பு

அ. 1920களின் பிற்பகுதியில் அறிமுகப்படுத்தப்பட்டது. √

ஆ. 1930களின் பிற்பகுதியில் அறிமுகப்படுத்தப்பட்டது.

இ. 1940களின் பிற்பகுதியில் அறிமுகப்படுத்தப்பட்டது

ஈ. 1950களின் பிற்பகுதியில் அறிமுகப்படுத்தப்பட்டது.

2) கூகிள் ஒளிவழி எழுத்துணரி பதிவிறக்கம் செய்ய பலர் விரும்ப வில்லை. ஏனெனில்

அ. மென்பொருள் தங்கள் கணினியைப் பாதிக்கும்

ஆ. மென்பொருள் அதிக விலை உள்ளது

இ. வேறு மென்பொருள்கள் கிடைக்கின்றன.

ஈ. மென்பொருள் தங்கள் கணினி வன்வட்டில் அதிக இடத்தை எடுக்கும்.√

3) கூகுள் ஒளிவழி எழுத்துணரி மென்பொருள்

அ. 248-க்கும் குறைவான உலக மொழிகளுக்குச் செயயல்படுகிறது.

ஆ. 248-க்கும் அதிகமான உலக மொழிகளுக்குச் செயயல்படுகிறது.√

இ. ஐரோப்பிய மொழிகளுக்கு மட்டுமே செயயல்படுகிறது.

ஈ. ஆங்கில மொழிக்கு மட்டுமே செயல்படுகிறது.

4) கூகிள் ஒளிவழி எழுத்துணரி

அ. டெசராக்டின் சார்புகளைப் பயன்படுத்தவில்லை.

ஆ. டெசராக்டின் சார்புகளைப் பயன்படுத்துகிறது.√

இ. வேறு சார்புகளைப் பயன்படுத்துகிறது.

இ. மேற்கண்ட மூன்றும் சரியில்லை

5) குர்முகி போன்ற ஒரு சில எழுத்துக்களுக்கு ஒளிவழி எழுத்துணரியின்

அ. வெளியீடு மிகவும் மோசமாக உள்ளது.√

ஆ. வெளியீடு மிகவும் நன்றாக உள்ளது.

இ. வெளியீடு சுமாராக உள்ளது.

ஈ. பயன்பாடு இல்லை.

6) கூகிள்

ஆ. டிசம்பர் 2015இல் கையெழுத்து ஒளிவழி எழுத்துணரியை வெளியிட்டது.

ஆ. டிசம்பர் 2016 இல் கையெழுத்து ஒளிவழி எழுத்துணரியை வெளியிட்டது.

இ. டிசம்பர் 2017 இல் கையெழுத்து ஒளிவழி எழுத்துணரியை வெளியிட்டது.

ஈ. டிசம்பர் 2018 இல் கையெழுத்து ஒளிவழி எழுத்துணரியை வெளியிட்டது. √

7) வினையூக்கி (catalistic) என்பது

அ. குறியீடு உள்ள பணிப்பாய்வு தானியங்கு தளமாகும்

ஆ. குறியீடு உள்ள பண்ப்பாய்வு மனித இயக்கத் தளமாகும்

இ. குறியீடு இல்லாத பணிப்பாய்வு தானியங்கு தளமாகும்√

ஈ. குறியீடு இல்லாத பணிப்பாய்வு மனித இயக்கத் தளமாகும்

8) ஒரு சாம்பல் அளவுப் படத்தை த்ரெஷோல்டிங் மூலம் கருப்பு மற்றும் வெள்ளைப்படமாக மாற்றும் முறை

அ. வளைவு திருத்தம்

ஆ. இயல்பாக்கம்

இ. சத்தம் குறைப்பு

ஈ. இருமையாக்கம்√

9) ---------- ஒரு சீரற்ற அளவிலான படத்தை ஒரு நிலையான மாற்றும் செயல்முறையாகும்.

அ. இருமையாக்கம்

ஆ. சத்தம்குறைப்பு

இ. இயல்பாக்கம் √

ஈ. வளைவு திருத்தம்

10) பண்புக்கூறு பிரித்தெடுக்கும் நுட்பங்களை

அ. இரண்டு வகுப்புகளாகப் பிரிக்கலாம்

ஆ. மூன்று வகுப்புகளாகப் பிரிக்கலாம்√

இ. நான்கு வகுப்புகளாகப் பிரிக்கலாம்

ஈ. ஐந்து வகுப்புகளாகப் பிரிக்கலாம்

11) எழுத்துக்குறி படத்தை இடம்சார் பண்புக்கூறுகளின் தொகுப்பாக மாற்ற ----- பயன்படுத்தப்படுகிறது

அ. அளவு மாறா பண்புக்கூறு மாற்றம் √

ஆ. கட்டமைப்பு நுட்பம்

இ. புள்ளிவிவர் நுட்பம்

ஈ. கலப்பின நுட்பம்

12) ----- பயன்படுத்தி எழுத்து பிக்சல்களின் உயரமும் அகலமும் கணக்கிடப்படுகின்றன.

அ. மண்டல அடிப்படையிலான முறையை

ஆ. குறியீட்டுப் பைனரி மாறுபாடு முறையை√

இ. கபார் சேனல் முறை

ஈ. பிலினியர் இண்டர்போலேஷன் டெக்கினிக்

13) ------ குறியிடப்பட்ட எழுத்துச் சரம் அகராதியைப் பயன்படுத்தி தமிழ் அச்சிடப்பட்ட எழுத்துக்களை உணரும் முறையை விவரித்தார்.

அ. நிரோமனி

ஆ. பிரோமனி

இ. சிரோமனி√

ஈ. குரோமனி

14) ---------- ஆஃப்லைன் தமிழ் எழுத்துக்களை உணர்தலுக்காக ஆக்டல் வரைபட (Octal graph conversion) மாற்றத்தைப் பயன்படுத்துகின்றனர்

அ.சுரேஷ் மற்றும் பிறர்

ஆ. சீதாலட்சுமி மற்றும் பிறர்

இ. அபர்ணா மற்றும்பிறர்

ஈ. ஜெகதீஷ் மற்றும் பிறர் √

15)

2. பொருத்துக (4 வினா-விடை தொகுப்பு அடங்கிய 10 வினாக்கள்)

1) செயல்முறைகளைப் பொருத்துக

அ. சாம்பல் அளவுப்படத்தை த்ர்ஷோல்டிங் மூலம் கருப்பு மற்றும் வெள்ளைப் படமாக மாற்றும் ஒருமுறை – வளைவு திருத்தம்

ஆ. எழுத்துக்குறி படத்தின் தீவிரத்தை மாற்றப் பயன்படுத்தப்படும் – இயல்பாக்கம்

இ. ஒரு சீரற்ற அளவிலான படத்தை ஒரு நிலையான அளவாக மாற்றும் செயல்முறை – சராசரி வடிப்பான்

ஈ. ஒற்றை பிக்சல் அகலப் படத்தைக் கையால் எழுதப்பட்ட எழுத்தை எளிதில் அடையாளம் காணும் – இருமையாக்கம்

3. சரியா தவறா (15 வினாக்கள்)

1) ஒளிவழி எழுத்துணரி என்பது ரியாலிட்டி உலகத்தையும் மெய்நிகர் வார்த்தையும் இணைப்பதற்கான வழிகளில் ஒன்றாகும்.

√சரி/தவறு

2)கூகிள் ஒளிவழி எழுத்துணரியைப் பயன்படுத்த தனிநபருக்குத் தேவையானது ஒன்றுமில்லை.

சரி/தவறு√

3) கூகிள் வழி எழுத்துணரியை பதிவிறக்கம் செய்ய பலர் விரும்பவில்லை, ஏனெனில் மென்பொருள் தங்கள் கணினி வன்வட்டில் அதிக இடத்தை எடுக்கும் என்று அவர்கள் அஞ்சுகிறார்கள்.

√சரி/தவறு

4) கூகிள் கிளவுட் பப்/சப் பல்வேறு பணிகளை வரிசைப் படுத்தவும் சரியான கிளவுட் செயல்பாடுகளைச் செயல்படுத்தவும் பயன்படுகிறது.

√சரி/தவறு

5) ஸ்கிரிப்ட் எழுத்து பேசும்மொழியைக் குறிக்கிறது மொழி எழுத்து முறையைக் குறிக்கிறது.

சரி/தவறு√

6) உரை அறிதல் படத்திலிருந்து உரையை அறிகிறது.

√சரி/தவறு

7) கூகிளின் எழுத்துணரி ஐரோப்பிய மொழிகளுக்கே செயல்படுகிறது.

சரி/தவறு√

8) குர்முகி போன்ற ஒரு சில ஸ்கிரிப்டுகளுக்கு எழுத்துணரியின் வெளியீடு மிகவும் மோசமாக உள்ளது.

சரி/தவறு√

9) வினையூக்கி (catalytic) என்பது ஒரு குறியீடு இல்லாத பணிப்பாய்வு ஆட்டோமேஷன் தளமாகும்.

√சரி/தவறு

10) வளைவு திருத்தம் என்பது ஒரு சாம்பல் அளவுப் படத்தை த்ரெஷோல்டிங் மூலம் கருப்பு மற்றும் வெள்ளைப் படமாக மாற்றும் முறையாகும்.

சரி/தவறு√

11)இயல்பாக்கம் என்பது ஒரு சீரற்ற அளவிலான படத்தை ஒரு நிலையான அளவாக மாற்றும் செயல்முறையாகும்.

√சரி/தவறு

12) பிரித்தல் என்பது ஒரு செயல்முறையாகும், இது ஆவணப் படங்களைக் கோடுகள் சொற்கள் மற்று எழுத்துக்களாக இணைக்கப் பயன்படுகின்றது.

சரி/தவறு √

13) எழுத்துக்குறி படத்தை இடம்சார் பண்புக்கூறுகளின் தொகுப்பாக மாற்ற, அளவு மாறா பண்புக்கூறு மாற்றம் பயன்படுத்தப்படுகின்றது.

√சரி/தவறு

14) பிலினியர் இண்டர்போலேஷன் டெக்னிக் பயன்படுத்தி அனைத்துப் படங்களும் வேறுபட்ட உயரம் மற்றும் அகலத்திற்கு அளவிடப்படுகின்றன.

சரி/தவறு√

15) கிடைமட்ட மற்றும் செங்குத்துக் கோடுகளைக் கண்டறிய ஹஃப் டிரான்ஸ்ஃபார்ம் பயன்படுகிறது.

√சரி/தவறு

15) கே-அருகிலுள்ள அண்டையர் அணுகுமுறை எழுத்துக்குறி தொகுப்புகளையும் சிறந்த துல்லியத்தையும் உணர வகைப்படுத்தியாகப் பயன்படுகின்றது.

√சரி/தவறு

4. ஒரு சொல்/சொற்றொடரில் விடைதருக. (15 வினாக்கள்)

1) டெசராக்ட் ஹெச்பி மூலம் கண்டறியப்பட்டது மற்றும் மேம்பாடு எப்பொழுது முதல் கூகிள் நிதியுதவி அளித்துள்ளது.

2)திசை அடையாளத்தின் செயல்பாடு என்ன?

3) ஒரு சாம்பல் அளவுப் படத்தை த்ரெஷோல்டிங் மூலம் கருப்பு மற்றும் வெள்ளைப் படமாக மாற்றும் ஒரு முறையின் பெயர் என்ன?

4) டிஜிட்டல் படங்கள் பல வகையான சத்தங்களுக்கு ஆளாகின்றன. அதை அகற்றும் செயல் முறையின் பெயர் என்ன?

5) ஒரு சீரற்ற அளவிலன படத்தை ஒரு நிலையான அளவாக மாற்றும் செயல்முறையின் பெயர் என்ன?

6) தேவையற்ற பிக்சல்களை அகற்றுவதற்கு எந்த வழிமுறை பயன்படுத்தப்படுகின்றது?

7) எந்த செயல்முறை ஆவணப் படங்களைக் கோடுகள், சொற்கள் மற்றும் எழுத்துக்களாகப் பிரிக்கப் பயன்படுகிறது?

8) பண்புக்கூறு பிரித்தெடுத்தலை எவ்வாறு மூன்று வகுப்புகளாகப் பிரிக்கலாம்?

9) எழுத்துக்குறி படத்தை இடம்சார் பண்புக்கூறுகளின் தொகுப்பாக மாற்ற எது பயன்படுகின்றது?

10) மண்டல அடிப்படையிலான முறையில் இயல்பாக்கப்பட்ட எழுத்துக்கள் எப்படி பிரிக்கப்படுகின்றன?

11) பிலின்யர் இண்டர்பொலேஷன் டெக்னிக் பயன்படுத்தி அனைத்து படங்களும் எவ்வாறு அளவிடப் படுகின்றன?

12) கிடைமட்ட மற்றும் செங்குத்துக் கோடுகளைக் கண்டறிய எது பயன்படுத்தப்படுகிறது.

13) உணரலின் முடிவு எந்த சமன்பாட்டைப் பயன்படுத்திப் பெறப்படுகிறது.

14) தமிழ் எழுத்துக்களை உணர்வதில் எந்த இரண்டு வெவ்வேறு அணுகுமுறைகளின் சோதனைகளை நிரஞ்சன் மற்றும் பிறர் ஒப்பிடுகின்றனர்.

15) தமிழ் எழுத்துக்களை உணரப் பிரபலமான மீள் பொருத்தம் வழிமுறையான டைனமிக் டைம் மடக்குதலை யார் பயன்படுதினர்?

5.ஒரு பத்தியில் விடை தருக. (பத்து வினாக்கள்)

1) இருமையாக்கம் என்பது பற்றி ஒரு பத்தியில் எழுதுக.

2) சத்தம் அகற்றுதல் பற்றி ஒரு பத்தியில் எழுதுக.

3)இயல்பாக்கம் பற்றி சுருக்கமாகக் கூறுக.

4) பிரித்தல்பற்றி ஒரு பத்தியில் கூறுக.

5) பண்புக்கூறுகள் பிரித்தெடுத்தல் பற்றி எழுதுக.

6) புள்ளிவிவர நுட்பம் பற்றிக் கூறுக.

7) வகைப்படுத்தல் பற்றி ஒரு பத்தியில் விளக்குக.

8) கையெழுத்து உணர்தலில் நிகழ்தகவு இல்லாத மாதிகள் பற்றி சுருக்கிக் கூறுக.

9) கையெழுத்து உணர்தலில் நரம்பியல் நெட்வொர்க்குகள் பற்றி விளக்குக.

10) கையெழுத்து உணர்தலில் சுரேஷ் மற்றும் பிறரின் பங்களிப்பு பற்றிக் கூறுக.

6. மூன்று பக்க அளவில் விடை தருக (5 வினாக்கள்)

1) கூகிள் ஒளிவழி எழுத்துணரியை எவ்வாறு பயன்படுத்துவது என்பது குறித்து கட்டுரை வரைக.

2. கூகிள் ஒளிவழி எழுத்துணரி டுடோரியல் பற்றி ஒரு கட்டுரை எழுத்துக.

2) ஒளிவழி எழுத்துணர்தலில் முன்செயலாக்கம் பற்றி ஒரு கட்டுரை வரைக.

3) ஒளிவழி எழுத்துணர்தலில் தொழில்நுட்பம் மற்றும் வழிமுறைகள் பற்றி விளக்குக.

4) ஒளிவழி எழுத்துணர்தலில் செய்யப்பட்ட தனி நபர்களின் முயற்சி குறித்து விளக்குக.